On the Computational Complexity of MCMC-based Estimators in Large Samples

Alexandre Belloni[*]

Alexandre Belloni
e-mail: `abn5@duke.edu`
Decision Sciences, Fuqua School of Business
Duke Univeristy
1 Towerview Drive
PO Box 90120
Durham, NC 27708-0120
http://www.duke.edu/∼abn5

and

Victor Chernozhukov[†]

Victor Chernozhukov
e-mail: `vchern@mit.edu`
Department of Economics
Operations Research Center
Massachusetts Institute of Technology
50 Memorial Drive
Room E52-371f
Cambridge , MA 02142
http://web.mit.edu/vchern/www

Abstract: In this paper we examine the implications of the statistical large sample theory for the computational complexity of Bayesian and quasi-Bayesian estimation carried out using Metropolis random walks. Our analysis is motivated by the Laplace-Bernstein-Von Mises central limit theorem, which states that in large samples the posterior or quasi-posterior approaches a normal density. Using the conditions required for the central limit theorem to hold, we establish polynomial bounds on the computational complexity of general Metropolis random walks methods in large samples. Our analysis covers cases where the underlying log-likelihood or extremum criterion function is possibly non-concave, discontinuous, and with increasing parameter dimension. However, the central limit theorem restricts the deviations from continuity and log-concavity of the log-likelihood or extremum criterion function in a very specific manner.

Under minimal assumptions required for the central limit theorem to hold under the increasing parameter dimension, we show that the Metropolis algorithm is theoretically efficient even for the canonical Gaussian walk which is studied in detail. Specifically, we show that the running time of

---

the algorithm in large samples is bounded in probability by a polynomial in the parameter dimension d, and, in particular, is of stochastic order d^2 in the leading cases after the burn-in period. We then give applications to exponential families, curved exponential families, and Z-estimation of increasing dimension.

AMS 2000 subject classifications: Primary, 65C05; secondary 65C60.

Keywords and phrases: Markov Chain Monte Carlo, Computational Complexity, Bayesian, Increasing Dimension.

1. Introduction

Markov Chain Monte Carlo (MCMC) algorithms have dramatically increased the use of Bayesian and quasi-Bayesian methods for practical estimation and inference. (See e.g. books of Casella and Robert [9], Chib [12], Geweke [18], Liu [35] for detailed treatments of the MCMC methods and their applications in various areas of statistics, econometrics, and biometrics.) Bayesian methods rely on a likelihood formulation, while quasi-Bayesian methods replace the likelihood with other criterion functions. This paper studies the computational complexity of MCMC algorithms (based on Metropolis random walks) as both the sample and parameter dimensions grow to infinity at the appropriate rates. The paper shows how and when the large sample asymptotics places sufficient restrictions on the likelihood and criterion functions that guarantee the efficient – that is, polynomial time – computational complexity of these algorithms. These results suggest that at least in large samples, Bayesian and quasi-Bayesian estimators can be computationally efficient alternatives to maximum likelihood and extremum estimators, most of all in cases where likelihoods and criterion functions are non-concave and possibly non-smooth in the parameters of interest.

To motivate our analysis, let us consider the Z-estimation problem, which is a basic method for estimating various kinds of structural models, especially in biometrics and econometrics. The idea behind this approach is to maximize some criterion function:

$$Q_n(\theta) = -\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n m(U_i, \theta) \right\|^2, \quad \theta \in \Theta \subset \mathbb{R}^d, \tag{1.1}$$

where U_i is a vector of random variables, and $m(U_i, \theta)$ is a vector of functions such that $E[m(U_i, \theta)] = 0$ at the true parameter value $\theta = \theta_0$. For example, in estimation of conditional α-quantile models with censoring and endogeneity, the functions take the form

$$m(U_i, \theta) = W(\alpha/p_i(\theta) - 1(Y_i \leq X_i\theta))Z_i. \tag{1.2}$$

Here $U_i = (Y_i, X_i, Z_i)$, Y_i is the response variable, X_i is a vector of regressors; in the censored regression models, Z_i is the same as X_i, and $p_i(\theta)$ is a weighting function that depends on the probability of censoring that depends on X_i and θ (see [34] for extensive motivation and details), and in the endogenous models,

Z_i is a vector of instrumental variables that affect the outcome variable Y_i only through X_i (see [11] for motivation and details), while $p_i(\theta) = 1$ for each i; the matrix W is some positive definite weighting matrix. Finally, the index $\alpha \in (0, 1)$ is the quantile index, and $X_i'\theta$ is the model for the α-th quantile function of the outcome Y_i.

In these quantile examples, the criterion function $Q_n(\theta)$ is highly discontinuous and non-concave, implying that the argmax estimator may be difficult or impossible to obtain. Figure 1 in Section 2 illustrates this example and similar examples where the argmax computation is intractable, at least when the parameter dimension d is high. In typical applications, the parameter dimension d is indeed high in relation to the sample size (see e.g. Koenker [32] for a relevant survey). Similar issues can also arise in M-estimation problems, where the extremum criterion function takes the form, $Q_n(\theta) = \sum_{i=1}^{n} m(U_i, \theta)$, where U_i is a vector of random variables, and $m(U_i, \theta)$ is a real-valued function, for example, the log-likelihood function of U_i or some other pseudo-log-likelihood function. Section 5 discusses several examples of this kind.

As an alternative to argmax estimation in both the Z- and M-estimation frameworks, consider the quasi-Bayesian estimator obtained by integration in place of optimization:

$$\widehat{\theta} = \frac{\displaystyle\int_{\Theta} \theta \exp\{Q_n(\theta)\}d\theta}{\displaystyle\int_{\Theta} \exp\{Q_n(\theta')\}d\theta'}. \tag{1.3}$$

This estimator may be recognized as a quasi-posterior mean of the quasi-posterior density $\pi_n(\theta) \propto \exp Q_n(\theta)$. (Of course, when Q_n is a log-likelihood, the term "quasi" becomes redundant.) This estimator is not affected by local discontinuities and non-concavities and is often much easier to compute in practice than the argmax estimator, particulary in the high-dimensional setting; see, for example, the discussion in Liu, Tian, and Wei [34] and Chernozhukov and Hong [11].

At this point, it is worth emphasizing that we will formally capture the high parameter dimension by using the framework of Huber [23], Portnoy [42], and others. In this framework, we have a sequence of models (rather than a fixed model) where the parameter dimension grows as the sample size grows, namely, $d \to \infty$ as $n \to \infty$, and we will carry out all of our analysis in this framework.

This paper will show that if the sample size n grows to infinity and the dimension of the problem d does not grow too quickly relative to the sample size, the quasi-posterior

$$\frac{\exp\{Q_n(\theta)\}}{\displaystyle\int_{\Theta} \exp\{Q_n(\theta')\}d\theta'} \tag{1.4}$$

will be approximately normal. This result in turn leads to the main claim: the estimator (1.3) can be computed using Markov Chain Monte Carlo in *polynomial time*, provided that the starting point is drawn from the approximate support

of the quasi-posterior (1.4). As is standard in the literature, we measure running time in the number of evaluations of the numerator of the quasi-posterior function (1.4) since this accounts for most of the computational burden.

In other words, when the central limit theorem (CLT) for the quasi-posterior holds, the estimator (1.3) is computationally tractable. The reason is that the CLT, in addition to implying the approximate normality and attractive estimation properties of the estimator $\widehat{\theta}$, bounds non-concavities and discontinuities of $Q_n(\theta)$ in a specific manner that implies that the computational time is polynomial in the parameter dimension d. In particular, in the leading cases the bound on the running time of the algorithm after the so-called burn-in period is $O_p(d^2)$. Thus, our main insight is to bring the structure implied by the CLT into the computational complexity analysis of the MCMC algorithm for computation of (1.3) and sampling from (1.4).

Our analysis of computational complexity builds on several fundamental papers studying the computational complexity of Metropolis procedures, especially Applegate and Kannan [2], Frieze, Kannan and Polson [16], Polson [41], Kannan, Lovász and Simonovits [29], Kannan and Li [28], Lovász and Simonovits [37], and Lovász and Vempala [38, 39, 40]. Many of our results and proofs rely upon and extend the mathematical tools previously developed in these works. We extend the complexity analysis of the previous literature, which has focused on the case of an arbitrary concave log-likelihood function, to the nonconcave and nonsmooth cases. The motivation is that, from a statistical point of view, in concave settings it is typically easier to compute a maximum likelihood or extremum estimate than a Bayesian or quasi-Bayesian estimate, so the latter do not necessarily have practical appeal. In contrast, when the log-likelihood or quasi-likelihood is either nonsmooth, nonconcave, or both, Bayesian and quasi-Bayesian estimates defined by integration are relatively attractive computationally, compared to maximum likelihood or extremum estimators defined by optimization.

Our analysis relies on statistical large sample theory. We invoke limit theorems for posteriors and quasi-posteriors for large samples as $n \to \infty$. These theorems are necessary to support our principal task – the analysis of the computational complexity under the restrictions of the CLT. As a preliminary step of our computational analysis, we state a CLT for quasi-posteriors and posteriors under parameters of increasing dimension, which extends the CLT previously derived in the literature for posteriors and quasi-posteriors for fixed dimensions. In particular, Laplace c. 1809, Blackwell [7], Bickel and Yahav [6], Ibragimov and Hasminskii [24], and Bunke and Milhaud [8] provided CLTs for posteriors. Blackwell [7], Liu, Tian, and Wei [34], and Chernozhukov and Hong [11] provided CLTs for quasi-posteriors formed using various non-likelihood criterion functions. In contrast to these previous results, we allow for increasing dimensions. Ghosal [20] previously derived a CLT for posteriors with increasing dimension for log-concave exponential families. We go beyond this canonical setup and establish the CLT for the non-log-concave and discontinuous cases. We also allow for general criterion functions to replace likelihood functions. This paper also illustrates the plausibility of the approach using exponential

families, curved exponential families, and Z-estimation problems. The curved families arise for example when the data must satisfy additional moment restrictions, as e.g. in Hansen and Singleton [21], Chamberlain [10], and Imbens [25]. Both the curved exponential families and Z-estimation problems typically fall outside the log-concave framework.

The rest of the paper is organized as follows. In Section 2, we establish a generalized version of the Central Limit Theorem for Bayesian and quasi-Bayesian estimators. This result may be seen as a generalization of the classical Bernstein-Von-Mises theorem, in that it allows the parameter dimension to grow as the sample size grows. In Section 2, we also formulate the main problem, which is to characterize the complexity of MCMC sampling and integration as a function of the key parameters that describe the deviations of the quasi-posterior from the normal density. Section 3 explores the structure set forth in Section 2 to find bounds on conductance and mixing time of the MCMC algorithm. Section 4 derives bounds on the integration time of the standard MCMC algorithm. Section 5 considers an application to a broad class of curved exponential families and Z-estimation problems, which have possibly non-concave and discontinuous criterion functions, and verifies that our results apply to this class of statistical models. Section 5 also verifies that the high-level conditions of Section 2 follow from the primitive conditions for these models.

Comment 1.1 (Notations.) *Throughout the paper, we follow the framework of high dimensional parameters introduced in Huber (1973). In this framework the parameter $\theta^{(n)}$ of the model, the parameter space $\Theta^{(n)}$, its dimension $d^{(n)}$, and all other properties of the model itself are indexed by the sample size n, and $d^{(n)} \to \infty$ as $n \to \infty$. However, following Huber's convention, we will omit the index and write, for example, θ, Θ, and d as abbreviations for $\theta^{(n)}$, $\Theta^{(n)}$, and $d^{(n)}$, and so on.*

2. The Setup and The Problem

Our analysis is motivated by the problems of estimation and inference in large samples under high dimension. We consider a "reduced-form" setup formulated in terms of parameters that characterize local deviations from the true parameter value. The local parameter λ describes contiguous deviations from the true parameter shifted by a first order approximation to an extremum estimator $\tilde{\theta}$. That is, for θ denoting a parameter vector, θ_0 the true value, and $s = \sqrt{n}(\tilde{\theta} - \theta_0)$ the normalized first order approximation of the extremum estimator, we define the local parameter λ as

$$\lambda = \sqrt{n}(\theta - \theta_0) - s.$$

The parameter space for θ is Θ, and the parameter space for λ is therefore $\Lambda = \sqrt{n}(\Theta - \theta_0) - s$.

The corresponding localized likelihood or localized criterion function is denoted by $\ell(\lambda)$. For example, suppose $L_n(\theta)$ is the original likelihood function in the likelihood framework or, more generally, $L_n(\theta)$ is $\exp\{Q_n(\theta)\}$ where $Q_n(\theta)$

is the criterion function in extremum framework, then

$$\ell(\lambda) = L_n(\theta_0 + (\lambda + s)/\sqrt{n})/L_n(\theta_0 + s/\sqrt{n}).$$

The assumptions below will be stated directly in terms of $\ell(\lambda)$. In Section 5, we further illustrate the connection between the localized set-up and the non-localized set-ups and provide more primitive conditions within the exponential family, curved exponential family, and Z-estimation framework.

Then, the posterior or quasi-posterior density for λ takes the form, implicitly indexed by the sample size n,

$$f(\lambda) = \frac{\ell(\lambda)}{\int_\Lambda \ell(\omega)d\omega}, \tag{2.5}$$

and we impose conditions that force the posterior to satisfy a CLT in the sense of approaching the normal density

$$\phi(\lambda) = \frac{1}{(2\pi)^{d/2} \det(J^{-1})^{1/2}} \exp\left(-\frac{1}{2}\lambda' J\lambda\right). \tag{2.6}$$

More formally, the following conditions are assumed to hold for $\ell(\lambda)$ as the sample size and parameter dimension grow to infinity:

$$n \to \infty \quad \text{and} \quad d \to \infty.$$

We call these conditions the "CLT conditions":

C.1 The local parameter λ belongs to the local parameter space $\Lambda \subset \mathbb{R}^d$. The vector s is a zero mean vector with variance Ω, whose eigenvalues are bounded above as $n \to \infty$, and $\Lambda = K \cup K^c$, where K is a closed ball $B(0, \|K\|)$ such that $\int_K f(\lambda)d\lambda \geq 1 - o_p(1)$ and $\int_K \phi(\lambda)d\lambda \geq 1 - o(1)$.

C.2 The lower semi-continuous posterior or quasi-posterior function $\ell(\lambda)$ approaches a quadratic form in logs, uniformly in K, i.e., there exist positive approximation errors ϵ_1 and ϵ_2 such that for every $\lambda \in K$,

$$\left| \ln \ell(\lambda) - \left(-\frac{1}{2}\lambda' J\lambda\right) \right| \leq \epsilon_1 + \epsilon_2 \cdot \lambda' J\lambda/2, \tag{2.7}$$

where J is a symmetric positive definite matrix with eigenvalues bounded away from zero and from above uniformly in the sample n. Also, we denote the ellipsoidal norm induced by J as $\|v\|_J := \|J^{1/2}v\|$.

C.3 The approximation errors ϵ_1 and ϵ_2 satisfy $\epsilon_1 = o_p(1)$, and $\epsilon_2 \cdot \|K\|_J^2 = o_p(1)$.

Comment 2.1 *We choose the support set $K = B(0, \|K\|)$, which is a ball of radius $\|K\| = \sup_{\lambda \in K} \|\lambda\|$, as follows. Under increasing dimension, the normal density is subject to a concentration of measure, namely that selecting $\|K\| \geq C \cdot \sqrt{d}$, for a sufficiently large constant C, is enough to contain the support of the standard normal vector. Indeed, let $Z \sim N(0, I_d)$, then $Pr(Z \notin$*
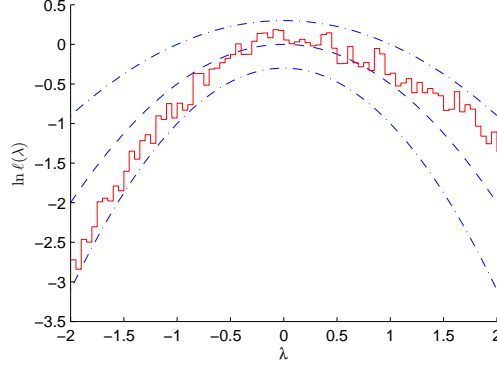
FIG 1. *This figure illustrates how* $\ln \ell(\lambda)$ *can deviate from* $\ln a(\lambda)$, *allowing for possible discontinuities in* $\ln \ell(\lambda)$.

$K) = Pr(\|Z\|^2 > C^2 d) \to 0$ *for* $C > 1$ *as* $d \to \infty$, *because* $\|Z\|^2/d \to_p 1$. *For the case where* $W \sim N(0, J^{-1}) = J^{-1/2}Z$, *we have that* $Pr(W \notin K) \leq Pr(\|Z\|/\sqrt{\lambda_{\min}} > \|K\|) \to 0$ *for* $\|K\| \geq C\sqrt{d/\lambda_{\min}}$ *for* $C > 1$ *as* $d \to \infty$, *where* $\lambda_{\min}$ *denotes the smallest eigenvalue of* J. *Moreover, since* $\|K\|_J = \lambda_{\max}\|K\|$, *where* $\lambda_{\max}$ *denotes the largest eigenvalue of* J, *we need to have that* $\|K\|_J > \sqrt{d\lambda_{\max}/\lambda_{\min}}$. *In view of condition C.3, this requires* $\epsilon_2 d\lambda_{\max}/\lambda_{\min} = o_p(1)$ *and hence* $\epsilon_2 d = o_p(1)$. *Thus, in some of the computations presented below, we will set*

$$\|K\| = C\sqrt{d/\lambda_{\min}} \ \text{and} \ \|K\|_J = C\sqrt{d\lambda_{max}/\lambda_{min}} \ \text{for} \ C > 1.$$

Finally, even though we make the assumption of bounded eigenvalues of J, *we will emphasize the dependence on the eigenvalues in most proofs and formal statements. This will allow us to see immediately the impact of changing this assumption.*

These conditions imply that

$$\ell(\lambda) = a(\lambda) \cdot m(\lambda)$$

over the approximate support set K, where

$$\ln a(\lambda) = -\frac{1}{2}\lambda' J\lambda, \tag{2.8}$$

$$-\epsilon_1 - \epsilon_2\lambda' J\lambda/2 \leq \ln m(\lambda) \leq \epsilon_1 + \epsilon_2\lambda' J\lambda/2. \tag{2.9}$$

Figure 1 illustrates the kinds of deviations of $\ln \ell(\lambda)$ from the quadratic curve captured by the parameters ϵ_1 and ϵ_2, and also shows the types of discontinuities and non-convexities permitted in our framework. Parameter ϵ_1 controls the size of local discontinuities and parameter ϵ_2 controls the global tilting away from the quadratic shape of the normal log-density.

Theorem 1 (Generalized CLT for Quasi-Posteriors) *Under conditions C.1-C.3, the quasi-posterior density (2.5) approaches the normal density (2.6) in the following sense:*

$$\int_\Lambda |f(\lambda) - \phi(\lambda)|d\lambda = o_p(1).$$

Theorem 1 is a simple preliminary result. However, the result is essential for defining the environment in which the main results of this paper – the computational complexity results – will be developed. The theorem shows that in large samples, provided that some regularity conditions hold, Bayesian and quasi-Bayesian inference have good large sample properties. The main part of the paper, namely Section 3, develops the *computational implications* of the CLT conditions. In particular, Section 3 shows that polynomial time computing of Bayesian and quasi-Bayesian estimators by MCMC is in fact implied by the CLT conditions. Therefore, the CLT conditions are essential for both good statistical properties of the posterior or quasi-posterior under increasing dimension, as shown in Theorem 1, and for good computational properties as shown in Section 3.

By allowing increasing dimension ($d \rightarrow \infty$) Theorem 1 extends the CLT previously derived in the literature for posteriors in the likelihood framework (Blackwell [7], Bickel and Yahav [6], Ibragimov and Hasminskii [24], Bunke and Milhaud [8], Ghosal [20], Shen [46]) and for quasi-posteriors in the general extremum framework, when the likelihood is replaced by general criterion functions (Blackwell [7], Liu, Tian, and Wei [34] and Chernozhukov and Hong [11]). The theorem also extends the results in Ghosal [20], who also considered increasing dimensions but focused his analysis to the exponential likelihood family framework. In contrast, Theorem 1 allows for non-exponential families and for quasi-posteriors in place of posteriors. Recall that quasi-posteriors result from using quasi-likelihoods and other criterion functions in place of the likelihood. This substantially expands the scope of the applications of the result. Importantly, Theorem 1 allows for non-smoothness and even discontinuities in the likelihood and criterion functions, which are pertinent in a number of applications listed in the introduction.

The Problem of the Paper. Our problem is to characterize the complexity of obtaining draws from $f(\lambda)$ and of Monte Carlo integration for computing

$$\int g(\lambda)f(\lambda)d\lambda,$$

where $f(\lambda)$ is restricted to the approximate support K. The procedure used to obtain the basic draws as well as to carry out Monte Carlo integration is a Metropolis random walk, which is a standard MCMC algorithm used in practice. The tasks are thus:

I. Characterize the complexity of sampling from $f(\lambda)$ as a function of $(d, n, \epsilon_1, \epsilon_2, K)$;

II. Characterize the complexity of calculating $\int g(\lambda)f(\lambda)d\lambda$ as a function of $(d, n, \epsilon_1, \epsilon_2, K)$;

III. Characterize the complexity of sampling from $f(\lambda)$ and performing integrations with $f(\lambda)$ in large samples as $d, n \to \infty$ by invoking the bounds on $(d, n, \epsilon_1, \epsilon_2, K)$ imposed by the CLT;

IV. Verify that the CLT conditions are applicable in a variety of statistical problems.

This paper formulates and solves this problem. Thus, the paper brings the CLT restrictions into the complexity analysis and develops complexity bounds for sampling and integrating from $f(\lambda)$ under these restrictions. These CLT restrictions, arising from the use of large sample theory and the imposition of certain regularity conditions, limit the behavior of $f(\lambda)$ over the approximate support set K in a specific manner that allows us to establish polynomial computing time for sampling and integration. Because the conditions for the CLT do not provide strong restrictions on the tail behavior of $f(\lambda)$ outside K other than C.1, our analysis of complexity is limited entirely to the approximate support set K defined in C.1-C.3.

By solving the above problem, this paper contributes to the recent literature on the computational complexity of Metropolis procedures. Early work was primarily concerned with the question of approximating the volume of high dimensional convex sets where uniform densities play a fundamental role (Lovász and Simonovits [37], Kannan, Lovász and Simonovits [29, 30]). Later, the approach was generalized for the cases where the log-likelihood is concave (Frieze, Kannan and Polson [16], Polson [41], and Lovász and Vempala [38, 39, 40]). However, under log-concavity the maximum likelihood or extremum estimators are usually preferred over Bayesian or quasi-Bayesian estimators from a computational point of view. Regarding cases in which log-concavity is absent, the settings in which there is great practical appeal for using Bayesian and quasi-Bayesian estimates, have received little treatment in the literature. One important exception is the paper of Applegate and Kannan [2], which covers nearly-log-concave but smooth densities using a discrete Metropolis algorithm. In contrast to Applegate and Kannan [2], our approach allows for both discontinuous and non-log-concave densities that are permitted to deviate from the normal density (not from an arbitrary log-concave density, like in Applegate and Kannan [2]) in a specific manner. The manner in which they deviate from the normal is motivated by the CLT and controlled by parameters ϵ_1 and ϵ_2, which are in turn restricted by the CLT conditions. Using the CLT restrictions also allows us to treat non-discrete sampling algorithms. In fact, it is known that the canonical Gaussian walk analyzed in Section 3.2.4 does not have good complexity properties (rapidly mixing) for arbitrary log-concave density functions, see Lovász and Vempala [40]. Nonetheless, the CLT conditions imply enough structure so that even a canonical Gaussian walk becomes in fact rapidly mixing. Moreover, the analysis is general in that it applies to any Metropolis chain, provided that it satisfies a simple geometric condition. We illustrate this condition with the canonical algorithm. This suggests that the same approach can be used to establish polynomial bounds for various more sophisticated schemes. Finally, as is standard in the literature, we assume that the starting point for the algorithm

occurs in the approximate support of the posterior. Indeed, the polynomial time bound that we derive applies only in this case because this is the domain where the CLT provides enough structure on the problem. Our analysis does not apply outside this domain.

3. The complexity of sampling using random walks

3.1. Set-Up and Main Result

In this section we bound the computational complexity of obtaining a draw from a random variable approximately distributed according to a density function f as defined in (2.5). (Section 4 builds upon these results to study the associated integration problem.) By invoking condition C.1, we restrict our attention entirely to the approximate support set K and the accuracy of sampling will be defined over this set. Consider a measurable space $(K, \mathcal{A})$. Our task is to draw a random variable according to a density function f restricted to K. This density induces a probability distribution on K defined by $Q(A) = \int_A f(x)dx / \int_K f(x)dx$ for any $A \in \mathcal{A}$. Asymptotically, it is well-known that random walks combined with a Metropolis filter are capable of performing such a task. Such random walks are characterized by an initial point u_0 and a *one-step* probability distribution, which depends on the current point, to generate the next candidate point of the random walk. The candidate point is accepted with a probability given by the Metropolis filter, which depends on the likelihood function ℓ, on the current and on the candidate point, and otherwise the random walk stays at the current point (see Casella and Robert [9] and Vempala [50] for details; Section 3.2.4 describes the canonical Gaussian random walk).

In the complexity analysis of this algorithm we are interested in bounding the number of steps of the random walk required to draw a random variable from Q with a given precision. Equivalently, we are interested in bounding the number of evaluations of the local likelihood function ℓ required for this purpose.

Next, following Lovász and Simonovits [37] and Vempala [50], we review definitions of concepts relevant for our analysis. Let $q(x|u)$ denote the probability density to generate a candidate point and $1_u(A)$ be the indicator function of the set A. For each $u \in K$ the one-step distribution P_u, the probability distribution after one step of the random walk starting from u, is defined as

$$P_u(A) = \int_{K \cap A} \min\left\{\frac{f(x)q(u|x)}{f(u)q(x|u)}, 1\right\} q(x|u)dx + (1 - p_u)1_u(A), \qquad (3.10)$$

where

$$p_u = \int_K \min\left\{\frac{f(x)q(u|x)}{f(u)q(x|u)}, 1\right\} q(x|u)dx \qquad (3.11)$$

is the probability of making a proper move, namely the move to $x \in K, x \neq u$, after one step of the chain from $u \in K$.

The triple $(K, \mathcal{A}, \{P_u : u \in K\})$, along with a starting distribution Q_0, defines a Markov chain in K. We denote by Q_t the probability distribution obtained

after t steps of the random walk. A distribution Q is called stationary on $(K, \mathcal{A})$ if for any $A \in \mathcal{A}$,

$$\int_K P_u(A)dQ(u) = Q(A). \tag{3.12}$$

Given the random walk described earlier, the unique stationary probability distribution Q is induced by the function f, $Q(A) = \int_A f(x)dx / \int_K f(x)dx$ for all $A \in \mathcal{A}$, see e.g. Casella and Roberts [9]. This is the main motivation for most of the MCMC studies found in the literature since it provides an asymptotic method to approximate the density of interest. As mentioned before, our goal is to properly quantify this convergence and for that we need to review additional concepts.

The ergodic flow of a set A with respect to a distribution Q is defined as

$$\Phi(A) = \int_A P_u(K \backslash A)dQ(u).$$

It measures the probability of the event $\{u \in A, u' \notin A\}$ where u is distributed according to Q and u' is distributed according to P_u; it captures the average flow of points leaving A in one step of the random walk. The measure Q is stationary if and only if $\Phi(A) = \Phi(K \backslash A)$ for all $A \in \mathcal{A}$ since

$$
\begin{aligned}
\Phi(A) &= \int_A P_u(K \setminus A)dQ(u) = \int_A (1 - P_u(A)) \, dQ(u) \\
&= Q(A) - \int_A P_u(A)dQ(u) = \int_K P_u(A)dQ(u) - \int_A P_u(A)dQ(u) \\
&= \Phi(K \setminus A).
\end{aligned}
$$

A Markov chain is said to be ergodic if $\Phi(A) > 0$ for every A with $0 < Q(A) < 1$, which is the case for the Markov chain induced by the random walk described earlier due to the assumptions on f, namely conditions C.1 and C.2.

Next we recall the concept of a conductance of a Markov chain, which plays a key role in the convergence analysis. Intuitively, a Markov chain will converge slowly to the steady state if there exists a set A in which the Markov chain stays "too long" relative to the measure of A or its complement $K \backslash A$. In order for a Markov chain to stay in A for a long time, the probability of stepping out of A with the random walk must be small, that is, the ergodic flow of A must be small relative to the measures of A and $K \backslash A$. The concept of conductance of a set A quantifies this notion:

$$\phi(A) = \frac{\Phi(A)}{\min\{Q(A), Q(K \backslash A)\}}, \quad 0 < Q(A) < 1.$$

The global conductance of the Markov chain is the minimum conductance over sets with positive measure

$$\phi = \inf_{A \in \mathcal{A}: 0 < Q(A) < 1} \phi(A). \tag{3.13}$$

Lovász and Simonovits [37] proved the connection between conductance and convergence for the continuous state space, and Jerome and Sinclair [26, 27] proved the connection for the discrete state space. We will extensively use Corolary 1.5 of Lovász and Simonovits [37], restated here as follows: Let Q_0 be M-warm with respect to the stationary distribution Q, namely

$$\sup_{A \in \mathcal{A}: Q(A) > 0} \frac{Q_0(A)}{Q(A)} = M, \qquad (3.14)$$

then, the total variation distance between the stationary distribution Q and the distribution Q_t, obtained after t steps of the Markov chain starting from Q_0, is bounded above by a function of global conductance ϕ and warmness parameter M:

$$\|Q_t - Q\|_{TV} = \sup_{A \in \mathcal{A}} |Q_t(A) - Q(A)| \le \sqrt{M} \left(1 - \frac{\phi^2}{2}\right)^t. \qquad (3.15)$$

Therefore, the global conductance ϕ determines the number of steps required to generate a random point whose distribution Q_t is within a specified distance of the target distribution Q. The conductance ϕ also bounds the autocovariance between consecutive elements of the Markov chain, which is important for analyzing the computational complexity of integration by MCMC; see Section 4 for a more detailed discussion. The warmness parameter M, which measures how the starting distribution Q_0 differs from the target distribution Q, also plays an important role in determining the quality of convergence of Q_t to Q. In what follows, we will calculate M explicitly for the canonical random walk.

The main result of this paper provides a lower bound for the global conductance of the Markov chain ϕ under the CLT conditions. In particular, we show that $1/\phi$ is bounded by a fixed polynomial in the dimension of the parameter space even for a canonical random walk considered in Section 3.2.4. In order to show this, we require the following geometric condition on the difference between the one-step distributions.

D.1 *There exist positive sequences h_n and c_n such that for every $u, v \in K$, $\|u - v\| \le h_n$ implies that*

$$\|P_u - P_v\|_{TV} < 1 - c_n.$$

D.2 *The sequences above can be taken to satisfy the following bounds*

$$\frac{1}{c_n \min\{h_n \sqrt{\lambda_{min}}, 1\}} = O_p(d).$$

Condition D.1 holds if at least a c_n-fraction of the probability distribution associated with P_u varies smoothly as the point u changes. Condition D.2 imposes a particular rate for the sequences. As shown in Theorem 2 below, the rates in Conditions D.1 and D.2 play an important role in delivering good, that is, polynomial time, computational complexity. We show in Section 3.2.4 that

Conditions D.1 and D.2 hold for the canonical Gaussian walk under Conditions C.1, C.2, and C.3. with

$$1/h_n = O_p(d) \ \text{ and } \ 1/c_n = O_p(1),$$

and λ_{min} bounded away from zero. Moreover, the rates in Condition D.2 appear to be sharp for the canonical Gaussian walk under our framework. It remains an important question whether different types of random walks could lead to better rates than those in Condition D.2 (see Vempala [50] for a relevant survey). Another interesting question is the establishment of lower bounds on the computational complexity of the type considered in Lovász [36].

Next we state the main result of the section.

Theorem 2 (Main Result on Complexity of Sampling) *Under Conditions C.1, C.2, and D.1, the global conductance of the induced Markov chain satisfies*

$$1/\phi = O \left(\frac{e^{2(\epsilon_1 + \epsilon_2 \|K\|_J^2/2)}}{c_n \min\{h_n \sqrt{\lambda_{\min}}, 1\}} \right). \tag{3.16}$$

In particular, a random walk satisfying these assumptions requires at most

$$N_\varepsilon = O_p \left(e^{4(\epsilon_1 + \epsilon_2 \|K\|_J^2/2)} \ \frac{\ln(M/\varepsilon)}{(c_n \min\{h_n \sqrt{\lambda_{min}}, 1\})^2} \right) \tag{3.17}$$

steps to achieve $\|Q_{N_\varepsilon} - Q\|_{TV} \leq \varepsilon$ where Q_0 is M-warm with respect to Q. Finally, if Conditions C.1, C.2, C.3, D.1 and D.2 hold, we have that

$$1/\phi = O_p(d)$$

and the number of steps N_ε is bounded by

$$O_p \left(d^2 \ \ln(M/\varepsilon) \right). \tag{3.18}$$

Thus, under the CLT conditions, Theorem 2 establishes the polynomial bound on the computing time, as stated in equation (3.18). Indeed, CLT conditions C.1 and C.2 first lead to the bound (3.17) and, then, condition C.3, which imposes $\epsilon_1 = o_p(1)$ and $\epsilon_2 \cdot \|K\|_J^2 = o_p(1)$, leads to the polynomial bound (3.18). It is also useful to note that, if the stated CLT conditions do not hold, the bound on the computing time needs not be polynomial: in particular, the first bound (3.17) is exponential in ϵ_1 and $\epsilon_2 \|K\|_J^2$. It is also useful to note that the approximate normality of posteriors and quasi-posteriors implied by the CLT conditions plays an important role in the proofs of this main result and of auxiliary lemmas. Therefore, the CLT conditions are essential for both (a) good statistical properties of the posterior or quasi-posterior under increasing dimension, as shown in Theorem 1 and (b) for good computational properties, as shown in Theorem 2. Thus, results (a) and (b) establish a clear link between the computational properties and the statistical environment.

The relevance of the particular random walk in bounding the conductance is captured through the parameters c_n and h_n defined in condition D.1. Theorem 2

shows that as long as we can take $1/c_n$ and $1/h_n$ to be bounded by a polynomial in the dimension of the parameter space d, we will obtain polynomial time guarantees for the sampling problem. In some cases, the warmness parameter M appearing in (3.18) can also be related to the particular random walk being used. This is the case in the canonical random walk discussed in detail in Section 3.2.4.

3.2. Proof of the Main Result

The proof of Theorem 2 relies on a new iso-perimetric inequality (Corollary 1) and a geometric property of the particular random walk (condition D.1). After the connection between the iso-perimetric inequality and the ergodic flow is established, the geometric property allows us to use the first result to bound the conductance from below. In what follows we provide an outline of the proof, auxiliary results, and, finally, the formal proof.

3.2.1. Outline of the Proof

The proof follows the arguments in Lovász and Simonovits [37] and Lovász and Vempala [38]. In order to bound the ergodic flow of $A \in \mathcal{A}$, consider the particular disjoint partition $K = \widetilde{S}_1 \cup \widetilde{S}_2 \cup \widetilde{S}_3$ where $\widetilde{S}_1 \subset A$, $\widetilde{S}_2 \subset K \setminus A$, and $\widetilde{S}_3$ consists of points in A or $K \setminus A$ for which the one-step probability of going to the other set is at least $c_n/2$ (to be defined later). Therefore we have

$$\begin{aligned}\Phi(A) \quad &= \int_A P_u(K \setminus A)dQ(u) = \tfrac{1}{2}\int_A P_u(K \setminus A)dQ(u) + \tfrac{1}{2}\int_{K \setminus A} P_u(A)dQ(u) \\ &\geq \tfrac{1}{2}\int_{\widetilde{S}_1} P_u(K \setminus A)dQ(u) + \tfrac{1}{2}\int_{\widetilde{S}_2} P_u(A)dQ(u) + \tfrac{c_n}{4}Q(\widetilde{S}_3).\end{aligned}$$

where the second equality holds because $\Phi(A) = \Phi(K \setminus A)$.

Since the first two terms could be arbitrarily small, the result will follow by bounding the last term from below. This will be achieved by a new iso-perimetric inequality tailored to the CLT framework and derived in Section 3.2.2. This result will provide a lower bound on $Q(\tilde{S}_3)$, which is increasing in the distance between $\widetilde{S}_1$ and $\widetilde{S}_2$.

Therefore, it remains to show that the distance between $\widetilde{S}_1$ and $\widetilde{S}_2$ is suitably bounded below. This follows from the geometric property stated in condition D.1. Given two points $u \in \widetilde{S}_1$ and $v \in \widetilde{S}_2$, we have $P_u(K \setminus A) \leq c_n/2$ and $P_v(A) \leq c_n/2$. Therefore, the total variation distance between their one-step distributions is bounded as

$$\|P_u - P_v\|_{TV} \geq |P_u(A) - P_v(A)| \geq 1 - c_n.$$

In such a case, condition D.1 implies that the distance $\|u - v\|$ is bounded from below by h_n. Since u and v are arbitrary points, the distance between sets $\widetilde{S}_1$ and $\widetilde{S}_2$ is bounded below by h_n.

This leads to a lower bound for the global conductance. After bounding the global conductance from below, Theorem 2 follows by invoking the conductance theorem of [37] restated in equation (3.15) and the CLT conditions.

3.2.2. An Iso-perimetric Inequality

We start by defining a notion of approximate log-concavity. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be log-β-concave if for every $\alpha \in [0,1]$, $x, y \in \mathbb{R}^d$, we have

$$f\left(\alpha x + (1 - \alpha)y\right) \geq \beta f(x)^\alpha f(y)^{1-\alpha}$$

for some $\beta \in (0,1]$, and f is said to be log-concave if β can be taken to be one. The class of log-β-concave functions is rather broad, including, for example, various non-smooth and discontinuous functions.

This concept is relevant under our CLT conditions C.1-C.3, since the relations (2.8) and (2.9) imposed by these conditions imply the following:

Lemma 1 *Over the set K, the functions $f(\lambda) := \ell(\lambda)/\int_\Lambda \ell(\lambda)d\lambda$ and $\ell(\lambda)$ can be written as the product of a Gaussian function, $e^{-\frac{1}{2}\lambda' J\lambda}$, and a log-$\beta$-concave function with parameter*

$$\beta = e^{-2(\epsilon_1 + \epsilon_2 \|K\|_J^2/2)}.$$

The representation of Lemma 1 gives us a convenient structure to establish the following iso-perimetric inequality.

Lemma 2 *Consider any measurable partition of the form $K = S_1 \cup S_2 \cup S_3$ such that the distance between S_1 and S_2 is at least t, i.e. $d(S_1, S_2) \geq t$. Let $Q(S) = \int_S f dx / \int_K f dx$. Then for any lower semi-continuous function $f(x) = e^{-\|x\|^2} m(x)$, where m is a log-β-concave function, we have*

$$Q(S_3) \geq \beta \frac{2te^{-t^2/4}}{\sqrt{\pi}} \min\{Q(S_1), Q(S_2)\}.$$

The iso-perimetric inequality of Lemma 2 states that if two subsets of K are far apart, the measure of the remaining subset of K should be comparable to the measure of at least one of the original subsets. This iso-perimetric inequality extends the iso-perimetric inequality in Kannan and Li [28]. The proof builds on their proof as well as on the ideas in Applegate and Kannan [2]. Unlike the inequality in Kannan and Li [28], Lemma 2 removes the smoothness assumptions on f, covering both non-log-concave and discontinuous cases.

The following corollary extends Lemma 2 to the case of an arbitrary covariance matrix J.

Corollary 1 (Iso-perimetric Inequality) *Consider any measurable partition of the form $K = S_1 \cup S_3 \cup S_2$ such that $d(S_1, S_2) \geq t$, and let $Q(S) = \int_S f dx / \int_K f dx$. Then, for any lower semi-continuous function $f(x) = e^{-\frac{1}{2}x' Jx} m(x)$, where m is a log-β-concave function and J is positive definite covariance matrix, we have*

$$Q(S_3) \geq \beta \ \sqrt{\lambda_{min}} t e^{-\lambda_{min} t^2/8} \ \sqrt{\frac{2}{\pi}} \min\{Q(S_1), \ Q(S_2)\},$$

where λ_{min} denotes the minimum eigenvalue of J.

3.2.3. Proof of Theorem 2

Fix an arbitrary set $A \in \mathcal{A}$ and denote by $A^c = K \setminus A$ the complement of A with respect to K. We will prove that

$$\Phi(A) \geq \frac{c_n}{4} \beta \sqrt{\frac{2}{\pi e}} \min \left\{ \frac{h_n}{2} \sqrt{\lambda_{min}}, 1 \right\} \min\{Q(A), Q(A^c)\}, \qquad (3.19)$$

where $\beta = e^{-2(\epsilon_1 + \epsilon_2 \|K\|_J^2/2)}$ is as defined in Lemma 1. This result implies the desired bound on the global conductance ϕ.

Consider the following auxiliary definitions:

$$\widetilde{S}_1 = \left\{ u \in A : P_u(A^c) < \frac{c_n}{2} \right\}, \widetilde{S}_2 = \left\{ v \in A^c : P_v(A) < \frac{c_n}{2} \right\}, \widetilde{S}_3 = K \setminus (\widetilde{S}_1 \cup \widetilde{S}_2).$$

In this case $Q(\widetilde{S}_1) \leq Q(A)/2$, we have

$$\begin{aligned}
\Phi(A) &= \int_A P_u(A^c) dQ(u) \geq \int_{A \setminus \widetilde{S}_1} P_u(A^c) dQ(u) \geq \int_{A \setminus \widetilde{S}_1} \frac{c_n}{2} dQ(u) \\
&\geq \frac{c_n}{2} Q(A \setminus \widetilde{S}_1) \geq \frac{c_n}{4} Q(A),
\end{aligned}$$

which immediately implies the inequality (3.19). In the case $Q(\widetilde{S}_2) \leq Q(A^c)/2$, we apply a similar argument.

In the remaining case $Q(\widetilde{S}_1) \geq Q(A)/2$ and $Q(\widetilde{S}_2) \geq Q(A^c)/2$, we proceed as follows. Since $\Phi(A) = \Phi(A^c)$ we have that

$$\begin{aligned}
\Phi(A) = \int_A P_u(A^c) dQ(u) &= \frac{1}{2} \int_A P_u(A^c) dQ(u) + \frac{1}{2} \int_{A^c} P_v(A) dQ(v) \\
&\geq \frac{1}{2} \int_{A \setminus \widetilde{S}_1} P_u(A^c) dQ(u) + \frac{1}{2} \int_{A^c \setminus \widetilde{S}_2} P_v(A) dQ(v) \\
&\geq \frac{1}{2} \int_{\widetilde{S}_3} \frac{c_n}{2} dQ(u) = \frac{c_n}{4} Q(\widetilde{S}_3),
\end{aligned}$$

where we used that $\widetilde{S}_3 = K \setminus (\widetilde{S}_1 \cup \widetilde{S}_2) = (A \setminus \widetilde{S}_1) \cup (A^c \setminus \widetilde{S}_2)$. Given the definitions of the sets $\widetilde{S}_1$ and $\widetilde{S}_2$, for every $u \in \widetilde{S}_1$ and $v \in \widetilde{S}_2$ we have

$$\|P_u - P_v\|_{TV} \geq P_u(A) - P_v(A) = 1 - P_u(A^c) - P_v(A) \geq 1 - c_n.$$

In such a case, by condition D.1, we have that $\|u - v\| > h_n$ for every $u \in \widetilde{S}_1$ and $v \in \widetilde{S}_2$. Thus, we can apply the iso-perimetric inequality of Corollary 1, with $d(\widetilde{S}_1, \widetilde{S}_2) \geq h_n$, to bound $Q(\widetilde{S}_3)$. We then obtain

$$\begin{aligned}
\int_A P_u(A^c) dQ(u) &\geq \max_{0 \leq t \leq h_n} \frac{c_n}{4} \beta \sqrt{\frac{2}{\pi}} \sqrt{\lambda_{\min}} \, t e^{-\frac{1}{8} \lambda_{min} t^2} \, \min\{Q(\widetilde{S}_1), Q(\widetilde{S}_2)\} \\
&\geq \frac{c_n}{4} \beta \sqrt{\frac{2}{\pi e}} \min \left\{ \frac{h_n}{2} \sqrt{\lambda_{min}}, 1 \right\} \min\{Q(A), Q(A^c)\}.
\end{aligned}$$

where we used the fact that $\max_{0 \leq t \leq h_n} \sqrt{\lambda_{\min}} t e^{-\frac{1}{8} \lambda_{min} t^2}$ is bounded below by $\min \left\{ h_n \sqrt{\lambda_{min}}, 2 \right\} e^{-1/2}$ and that $\min\{Q(\widetilde{S}_1), Q(\widetilde{S}_2)\} \geq \min\{Q(A), Q(A^c)\}/2$. Thus, the inequality (3.19) and the lower bound on conductance (3.16) follow.

The bound (3.17) on the number of steps of the Markov Chain follows from the lower bound on conductance (3.16) and the conductance theorem of [37] restated in equation (3.15). The remaining results in Theorem 2 follow by invoking the CLT conditions.∎

3.2.4. The case of the Gaussian random walk

In order to provide a concrete example of our complexity bounds, we consider the canonical random walk induced by a Gaussian distribution. Such a random walk is completely characterized by an initial point u_0, a fixed standard deviation $\sigma > 0$, and its *one-step* move. The latter is defined by the procedure of drawing a point y from a Gaussian distribution centered at the current point u with covariance matrix $\sigma^2 I$, and then if $y \in K$ moving to y with probability $\min\{f(y)/f(u), 1\} = \min\{\ell(y)/\ell(u), 1\}$, and otherwise staying at u.

We start with the following auxiliary result.

Lemma 3 *Let $a : \mathbb{R}^n \to \mathbb{R}$ be a function such that $\ln a$ is Lipschitz with constant L over a compact set K. Then, for every $u \in K$ and $r > 0$,*

$$\inf_{y \in B(u,r) \cap K} [a(y)/a(u)] \geq e^{-Lr}.$$

Given the ball $K = B(0, \|K\|)$, we can bound the Lipschitz constant of the function $-\lambda' J\lambda/2$ by

$$L = \sup_{\lambda \in K} \|J\lambda\| = \lambda_{max}\|K\|. \tag{3.20}$$

We define the parameter σ of the Gaussian random walk as

$$\sigma = \min\left\{ \frac{1}{4\sqrt{d}L}, \frac{\|K\|}{120d} \right\}. \tag{3.21}$$

Using (3.20) and that $\|K\| > \sqrt{d/\lambda_{\min}}$ it follows that

$$\sigma \geq \frac{1}{120\lambda_{max}\sqrt{d}\|K\|}. \tag{3.22}$$

In order to apply Theorem 2 we rely on σ being defined in (3.21) as a function of the relevant theoretical quantities. More practical choices of the parameter, as in Robert and Rosenthal [44] and Gelman, Roberts and Gilks [17], suggest that we tune the parameter to ensure a particular average acceptance rate for the steps of the Markov Chain. These cases are exactly the cases covered by our (theoretical) choice of σ (of course, different constant acceptance rates lead to different constants in the proof of the theorem). Moreover, a different choice of covariance matrix for the auxiliary Gaussian distribution can lead to improvements in practice but, under the assumptions on the matrix J, does not affect the overall dependence on the dimension d, which is our focus here.

Next we verify conditions D.1 and D.2 for the Gaussian random walk. Although this approach follows that in Lovász and Vempala [38, 39, 40], there are two important differences which call for a new proof. First, we no longer rely on the log-concavity of f. Second, we use a different random walk.

Lemma 4 *Let $u, v \in K := B(0, \|K\|)$, suppose that $\sigma \leq \min\{\frac{1}{4\sqrt{d}L}, \frac{\|K\|}{120d}\}$, and $\|u - v\| < \frac{\sigma}{8}$, where L is the Lipschitz constant specified in equation (3.20). Under conditions C.1-C.2, we have for $\beta = e^{-2(\epsilon_1 + \epsilon_2 \|K\|_J^2 / 2)}$ that*

$$\|P_u - P_v\|_{TV} \leq 1 - \frac{\beta}{3e}.$$

Comment 3.1 *Therefore, the Gaussian random walk satisfies condition D.1 with*

$$c_n = \frac{\beta}{3e} \quad and \quad h_n = \frac{\sigma}{8}. \tag{3.23}$$

Under the CLT framework, i.e. conditions C.1, C.2, and C.3, we have that c_n and h_n as defined in (3.23) satisfy condition D.2 with

$$1/h_n = O_p(d) \ and \ 1/c_n = O_p(1),$$

and λ_{min} bounded away from zero.

By applying Theorem 2 to the Gaussian random walk, the conductance bound (3.16) becomes

$$1/\phi = O\left(\frac{\lambda_{max}}{\lambda_{min}} \ d \ e^{2(\epsilon_1 + \epsilon_2 \|K\|_J / 2)}\right) = O_p(d)$$

and the bound on the number of steps N_ε in (3.17) becomes

$$O_p\big(d^2 \ln(M/\varepsilon)\big). \tag{3.24}$$

Next we discuss and bound the dependence on M, the "distance" of the initial distribution Q_0 from the stationary distribution Q as defined in (3.14). A natural candidate for a starting distribution Q_0 is the one-step distribution conditional on a proper move from an arbitrary point $u \in K$. Thus,

$$Q_0(A) = p_u^{-1} \cdot \int_{K \cap A} \min\left\{\frac{f(x)q(u|x)}{f(u)q(x|u)}, 1\right\} q(x|u)dx,$$

where

$$p_u = \int_K \min\left\{\frac{f(x)q(u|x)}{f(u)q(x|u)}, 1\right\} q(x|u)dx$$

is the probability of a proper move, namely the move to $x \in K, x \neq u$, after one step of the chain from $u \in K$. We emphasize that, in general, such choice of Q_0 could lead to values of M that are arbitrary large. In fact, this could happen even in the case of the stationary density being a uniform distribution on a convex set (see Lovász and Vempala [40]). However, this is not the case under the CLT framework as shown by the following lemma.

Lemma 5 *Suppose conditions C.1-C.2 hold, then for $\beta = e^{-2(\epsilon_1 + \epsilon_2 \|K\|_J^2/2)}$ we have that with a probability $p_u \geq \beta/(3e)$ the random walk makes a proper move. Moreover, let $u \in K$ and Q_0 be the associated one-step distribution conditional on performing a proper move starting from u, then Q_0 is M-warm with respect to Q, where*

$$\ln M = O(d\ln(\|K\|_J^2) + \|K\|_J^2 + \epsilon_1 + \epsilon_2\|K\|_J^2).$$

Under conditions $\epsilon_1 = o_p(1)$, $\epsilon_2\|K\|_J = o_p(1)$, and $\|K\|_J = O(\sqrt{d})$ we have

$$\ln M = O_p(d\ln d) \ \text{and} \ p_u \geq 1/(3e) + o_p(1).$$

Comment 3.2 (Overall Complexity for Gaussian Walk) *The combination of this result with relation (3.24), which was derived from Theorem 2, yields the overall (burn-in plus post burn-in) running time*

$$O_p(d^3 \ln d).$$

4. The complexity of Monte Carlo integration

This section considers our second task of interest – that of computing a high dimensional integral of a bounded real valued function g:

$$\mu_g = \int_K g(\lambda)dQ(\lambda). \tag{4.25}$$

Theorem 2 showed that the CLT conditions provide enough structure to bound the conductance of the Markov chain associated with a particular random walk. Below we also show how the conductance and CLT-based bounds on conductance impact the computational complexity of calculating (4.25) via standard schemes (long run, multiple runs, and subsampling). These new characterizations complement the previous well-known characterizations of the error in estimating (4.25) in terms of the covariance functions of the underlying chain (Geyer [19], Casella and Roberts [9], and Fishman [15]).

In what follows, a random variable λ^t is distributed according to Q_t, the probability measure obtained after iterating the chain t times, beginning from a starting measure Q_0. The chain $\lambda^t, t = 0, 1, ...$ has the stationary distribution Q. Accordingly, a standard estimate of (4.25), called the long-run (lr) average, takes the form

$$\widehat{\mu}_g = \frac{1}{N}\sum_{i=B}^{B+N} g(\lambda^i), \tag{4.26}$$

discarding the first B draws, the burn-in sample, and using subsequent N draws of the Markov chain.

The dependent nature of the chain increases the number of post-burn-in draws N needed to achieve a desired precision compared to the infeasible case of independent draws from Q. It turns out that, as in the preceding analysis,

the conductance of the Markov chain is crucial for determining the appropriate N.

The starting point of our analysis is a central limit theorem for reversible Markov chains due to Kipnis and Varadhan [31]: Consider a reversible Markov chain on K with a stationary distribution Q. The lag k autocovariance of the stationary time series $g(\lambda^i), i = 1, 2, ...,$ obtained by starting the Markov chain with the stationary distribution Q is defined as

$$\gamma_k = \text{Cov}_Q\left(g(\lambda^i), g(\lambda^{i+k})\right).$$

Then, for a stationary, irreducible, reversible Markov chain,

$$NE[(\widehat{\mu}_g - \mu_g)^2] \to \sigma_g^2 = \sum_{k=-\infty}^{+\infty} \gamma_k, \tag{4.27}$$

almost surely. If σ_g^2 is finite, then

$$\sqrt{N}(\widehat{\mu}_g - \mu_g) \to_d N(0, \sigma_g^2). \tag{4.28}$$

In our case, γ_0 is finite since g is bounded. Let us recall a result, which is due to Lovász and Simonovits [37], and which states that σ_g^2 can be bounded using the global conductance ϕ of a stationary, irreducible, reversible Markov chain: Let g be a square integrable function with respect to the stationary measure Q, then

$$|\gamma_k| \leq \left(1 - \frac{\phi^2}{2}\right)^{|k|} \gamma_0 \text{ and } \sigma_g^2 \leq \gamma_0\left(\frac{4}{\phi^2}\right). \tag{4.29}$$

We will use these conductance-based bounds to obtain bounds on the complexity of integration under the CLT conditions.

There exist other methods for constructing the sequence of draws in constructing estimators of the type (4.26); we refer to Geyer [19] for a detailed discussion. In addition to the long run (lr) method, we also consider the subsample (ss) and multi-start (ms) methods. Denote the number of post burn-in draws corresponding to each method as N_{lr}, N_{ss}, and N_{ms}. As mentioned above, the long run method consists of generating the first point using the starting distribution Q_0 and, after the burn-in period, selecting the N_{lr} subsequent points to compute the sample average. The subsample method also uses only one sample path, but the N_{ss} draws used in the sample average are spaced out by S steps of the chain. Finally, the multi-start method uses N_{ms} different sample paths, initializing each one independently from the starting probability distribution Q_0 and picking the last draw in each sample path after the burn-in period to compute the average. Thus, all estimators discussed above take the form

$$\widehat{\mu}_g = \frac{1}{N} \sum_{i=1}^{N} g(\lambda^{i,B})$$

with the underlying sequence $\lambda^{1,B}, \lambda^{2,B}, ..., \lambda^{N,B}$ produced as follows:

- for lr, $\lambda^{i,B} = \lambda^{i+B}$, where B is the burn-in period,
- for ss, $\lambda^{i,B} = \lambda^{iS+B}$, where S is the number of draws being skipped,
- for ms, $\lambda^{i,B}$ are i.i.d. draws from Q_B, that is, $\lambda^{i,B} \sim \lambda^B$ for every i.

There is a final issue that must be addressed. Both the central limit theorem of [31], restated in equations (4.27) and (4.28) and the conductance-based bound of [37] on covariances restated in equation (4.29) require that the initial point be drawn from the stationary distribution Q. However, we are starting the chain from some other distribution Q_0, and in order to apply these results we need to first run the chain for sufficiently many steps B, to bring the distribution of the draws Q_B close to Q in total variation metric. This is what we call the burn-in period. However, even after the burn-in period there is still a discrepancy between Q and Q_B, which should be taken into account. But once Q_B is close to Q, we can use the results on complexity of integration where sampling starts with Q to bound the complexity of integration where sampling starts with Q_B, where the bound depends on the discrepancy between Q_B and Q. Thus, our computational complexity calculations take into account all of the following three facts: (i) we are starting with a distribution Q_0 that is M-warm with respect to Q, (ii) from Q_0 we are making B steps with the chain in the burn-in period to obtain Q_B such that $\|Q_B - Q\|_{TV}$ is sufficiently small, and (iii) we are only using draws after the burn-in period to approximate the integral.

We use the mean square error as the measure of closeness for a consistent estimator:

$$MSE(\widehat{\mu}_g) = E\left[(\widehat{\mu}_g - \mu_g)^2\right].$$

Theorem 3 (Complexity of Integration) *Let Q_0 be M-warm with respect to Q, and let $\bar{g} := \sup_{\lambda \in K} |g(\lambda)|$. In order to obtain*

$$MSE(\widehat{\mu}_g) < \varepsilon$$

it is sufficient to use the following lengths of the burn-in sample, B, and post-burn-in samples, N_{lr}, N_{ss}, N_{ms}:

$$B = \left(\frac{2}{\phi^2}\right) \ln \left(\frac{24\sqrt{M}\bar{g}^2}{\varepsilon}\right)$$

and

$$N_{lr} = \frac{\gamma_0}{\varepsilon} \frac{6}{\phi^2}, \quad N_{ss} = \frac{3\gamma_0}{\varepsilon} \ \text{with} \ S = \frac{2}{\phi^2} \ln \left(\frac{6\gamma_0}{\varepsilon}\right), \quad N_{ms} = \frac{2\gamma_0}{3\varepsilon}.$$

The overall complexities of the lr, ss, and ms methods are thus $B+N_{lr}$, $B+SN_{ss}$, and $B \times N_{ms}$.

For convenience, Table 1 tabulates the bounds for the three different schemes. Note that the dependence on M and $\bar{g}$ is only via log terms. Although the optimal choice of the method depends on the particular values of the constants, when $\varepsilon \searrow 0$, the long-run algorithm has the smallest (best) bound, while the the

TABLE 1
Burn-in and Post Burn-in Bounds on the Complexity of Integration of a Bounded Function via Conductance

Method	Quantities	Complexity
Long Run	$B + N_{lr}$	$\frac{2}{\phi^2}\left(\ln\left(\frac{24\sqrt{M}\bar{g}^2}{\varepsilon}\right)\right) + \frac{2}{\phi^2}\left(\frac{3\gamma_0}{\varepsilon}\right)$
Subsample	$B + N_{ss} \cdot S$	$\frac{2}{\phi^2}\left(\ln\left(\frac{24\sqrt{M}\bar{g}^2}{\varepsilon}\right)\right) + \frac{2}{\phi^2}\left(\frac{3\gamma_0}{\varepsilon}\ln\left(\frac{24\gamma_0}{\varepsilon}\right)\right)$
Multi-start	$B \times N_{ms}$	$\frac{2}{\phi^2}\left(\ln\left(\frac{24\sqrt{M}\bar{g}^2}{\varepsilon}\right)\right) \times \frac{2\gamma_0}{3\varepsilon}$

TABLE 2
Burn-in and Post Burn-in Bounds on the Complexity of Integration of a Bounded Function using the Gaussian random walk under the CLT framework with
$\|K\|_J = O(\sqrt{d}), \epsilon_1 = o_p(1), \epsilon_2\|K\|_J^2 = o_p(1),$ *and* $\bar{g} = O(1).$

Method	Burn-in Complexity	Post-burn-in Complexity
Long Run	$O_p(d^3 \ln d \cdot \ln \varepsilon^{-1})$	$+ O_p(d^2 \cdot \varepsilon^{-1})$
Subsample	$O_p(d^3 \ln d \cdot \ln \varepsilon^{-1})$	$+ O_p(d^2 \cdot \varepsilon^{-1} \cdot \ln \varepsilon^{-1})$
Multi-start	$O_p(d^3 \ln d \cdot \ln \varepsilon^{-1})$	$\times O_p(\varepsilon^{-1})$

multi-start algorithm has the largest (worst) bound on the number of iterations. Table 2 presents the computational complexities implied by the CLT conditions, namely

$$\|K\|_J = O(\sqrt{d}), \epsilon_1 = o_p(1), \text{ and } \epsilon_2\|K\|_J^2 = o_p(1),$$

and the Gaussian random walk studied in Section 3.2.4. The table assumes γ_0 and $\bar{g}$ are constant, though it is straightforward to tabulate the results for the case where γ_0 and $\bar{g}$ grow at polynomial speed with d. Finally, note that the bounds apply under a slightly weaker condition than the CLT requires, namely that $\epsilon_1 = O_p(1)$ and $\epsilon_2\|K\|_J^2 = O_p(1)$.

5. Applications

In this section we verify that the CLT conditions and the analysis apply to a variety of statistical problems. In particular, we focus on the MCMC estimator (1.3) as an alternative to M- and Z-estimators. Here our goal is to derive the high-level conditions C1-C3 from appropriate primitive conditions, and thus show the efficient computational complexity of the MCMC estimator.

5.1. M-Estimation

We present two examples in M-estimation. We begin with the canonical log-concave cases within the exponential family. Then we drop the concavity and smoothness assumptions to illustrate the full applicability of the approach developed in this paper.

5.1.1. Exponential Family

Exponential families play a very important role in statistical estimation, cf. Lehmann and Casella [33], especially in high-dimensional contexts; see Portnoy [42], Ghosal [20], and Stone et al. [47]. For example, the high-dimensional situations arise in modern data sets in technometric and econometric applications. Moreover, exponential familes have excellent approximation properties and are useful for approximation of densities that are not necessarily of the exponential form; see Stone et al. [47].

We base our discussion on the asymptotic analysis of Ghosal [20]. In order to simplify the exposition, we invoke the more canonical conditions similar to those given in Portnoy [42]. Moreover, we assume that these conditions, numbered as E.1 to E.4, hold uniformly in the sample size n.

E.1 Let $X_1, \ldots, X_n$ be iid observations from a d-dimensional canonical exponential family with density

$$h(x; \theta) = \exp\left(x'\theta - \psi(\theta)\right),$$

where $\theta \in \Theta$ is an open subset of $\mathbb{R}^d$, and $d \to \infty$ as $n \to \infty$. Fix a sequence of parameter points $\theta_0 \in \Theta$. Set $\mu = \psi'(\theta_0)$ and $J = \psi''(\theta_0)$, the mean and covariance of the observations, respectively. Following Portnoy [42], we implicitly re-parameterize the problem, so that the Fisher information matrix $J = I$.

For a given prior π on Θ, the posterior density of θ over Θ conditioned on the data takes the form

$$\pi_n(\theta) \propto \pi(\theta) \cdot \prod_{i=1}^{n} h(X_i; \theta) = \pi(\theta) \cdot \exp\left(n\bar{X}'\theta - n\psi(\theta)\right),$$

where $\bar{X} = \sum_{i=1}^{n} X_i / n$ is the empirical mean of the data.

We associate every point θ in the parameter space Θ with a local parameter $\lambda \in \Lambda = \sqrt{n}(\Theta - \theta) - s$, where

$$\lambda = \sqrt{n}(\theta - \theta_0) - s,$$

and $s = \sqrt{n}(\bar{x} - \mu)$ is a first order approximation to the normalized maximum likelihood/extremum estimate. By design, we have that $E[s] = 0$ and $E[ss'] = I_d$. Moreover, by Chebyshev's inequality, the norm of s can be bounded in

probability, $\|s\| = O_p(\sqrt{d})$. Finally, the posterior density of λ over $\Lambda = \sqrt{n}(\Theta - \theta_0) - s$ is given by $f(\lambda) = \frac{\ell(\lambda)}{\int_\Lambda \ell(\lambda)d\lambda}$, where

$$
\begin{aligned}
\ell(\lambda) = \exp & \left(\bar{X}'\sqrt{n}\lambda - n\psi\left(\theta_0 + \frac{\lambda+s}{\sqrt{n}}\right) + n\psi\left(\theta_0 + \frac{s}{\sqrt{n}}\right) \right) \\
& \times \pi\left(\theta_0 + \frac{\lambda+s}{\sqrt{n}}\right) / \pi\left(\theta_0 + \frac{s}{\sqrt{n}}\right).
\end{aligned}
\tag{5.30}
$$

We impose the following regularity conditions, following Ghosal [20] and Portnoy [42]:

E.2 Consider the following quantities associated with higher moments in a neighborhood of the true parameter θ_0, uniformly in n :

$$
B_{1n}(c) := \sup_{\theta,\eta}\{E_\theta|\eta'(x_i - \mu)|^3 : \eta \in S^d, \|\theta - \theta_0\|^2 \le cd/n\},
$$
$$
B_{2n}(c) := \sup_{\theta,\eta}\{E_\theta|\eta'(x_i - \mu)|^4 : \eta \in S^d, \|\theta - \theta_0\|^2 \le cd/n\}.
$$

where $S^d = \{\eta \in \mathbb{R}^d : \|\eta\| = 1\}$. There are $p > 0$ and $c_0 > 0$ such that $B_{1n}(c) < c_0 + c^p$ and $B_{2n}(c) < c_0 + c^p$ for all $c > 0$ and all n.

E.3 The prior density π is proper and satisfies a positivity requirement at the true parameter

$$
\sup_{\theta \in \Theta} \ln\left[\pi(\theta)/\pi(\theta_0)\right] = O(d)
$$

where θ_0 is the true parameter. Moreover, the prior π also satisfies the following local Lipschitz condition

$$
|\ln \pi(\theta) - \ln \pi(\theta_0)| \le V(c)\sqrt{d}\|\theta - \theta_0\|
$$

for all θ such that $\|\theta - \theta_0\|^2 \le cd/n$, and some $V(c)$ such that $V(c) < c_0 + c^p$, with the latter holding for all $c > 0$.

E.4 The parameter dimension d grows at the rate such that $d^3/n \to 0$.

Condition E.2 strengthens an analogous condition of Ghosal [20], and implies an analogous assumption by Portnoy [42]. Condition E.3 is similar to the condition on the prior in Ghosal [20]. For further discussion of this condition, see [4]. Condition E.4 states that the parameter dimension should not grow too quickly relative to the sample size.

Theorem 4 *Conditions E.1-E.4 imply conditions C.1-C.3 with $\|K\| = C\sqrt{d}$ for some $C > 1$.*

Comment 5.1 *Combining Theorems 1 and 4, we have the asymptotic normality of the posterior,*
$$
\int_\Lambda |f(\lambda) - \phi(\lambda)|d\lambda = o_p(1).
$$

Furthermore, we can apply Theorem 2 to the posterior density f to bound the convergence time (number of steps) of the Metropolis walk needed to obtain a draw from f (with a fixed level of accuracy): The convergence time is at most

$$O_p(d^2)$$

after the burn-in period; together with the burn-in, the convergence time is

$$O_p(d^3 \ln d).$$

Finally, the integration bounds stated in the previous section also apply to the posterior f.

5.1.2. Curved Exponential Family

Next we consider the case of a d-dimensional curved exponential family. The curved family is general enough to allow for non-concavities and even non-smoothness in the log-likelihood function, which the canonical exponential family did not allow for. We assume that the following conditions, numbered as NE.1 to NE.4, hold uniformly in the sample size n, in addition to the previous conditions E.1 to E.4.

NE.1 Let $X_1, \ldots, X_n$ be iid observations from a d-dimensional curved exponential family with density

$$h(x; \theta) = \exp\left(x'\theta(\eta) - \psi(\theta(\eta))\right).$$

The parameter of interest is η, whose true value η_0 lies in the interior of a convex compact set $\Psi \subset \mathbb{R}^{d_1}$. The true value of θ, induced by η_0 is given by $\theta_0 = \theta(\eta_0)$. The mapping $\eta \mapsto \theta(\eta)$ takes values from $\mathbb{R}^{d_1}$ to $\mathbb{R}^d$ where $c \cdot d \leq d_1 \leq d$, for some $c > 0$. Finally, $d \to \infty$ as $n \to \infty$.

NE.2 True value η_0 is the unique solution to the system $\theta(\eta) = \theta_0$, and we have that $\|\theta(\eta) - \theta(\eta_0)\| \geq \epsilon_0 \|\eta - \eta_0\|$ for some $\epsilon_0 > 0$ and all $\eta \in \Psi$.

Thus, the parameter θ corresponds to a high-dimensional linear parametrization of the log-density, and η describes the lower-dimensional parametrization of the log-density. There are many classical examples of curved exponential families; see for example Efron [14], Lehmann and Casella [33], and Bandorff-Nielsen [3]. An example of the condition that puts a curved structure onto an exponential family is a moment restriction of the type:

$$\int m(x, \alpha) h(x, \theta) dx = 0.$$

This condition restricts θ to lie on a curve that can be parameterized as $\{\theta(\eta), \eta \in \Psi\}$, where the parameter $\eta = (\alpha, \beta)$ contains the component α as well as other components β. In econometric applications, moment restrictions often represent Euler equations that result from the data x being an outcome of an optimization
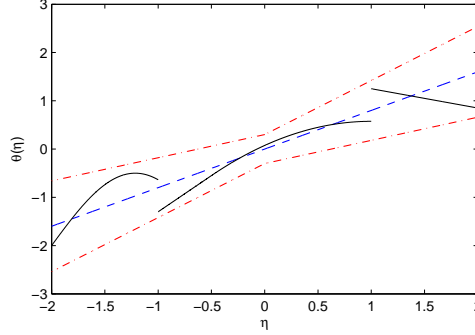
FIG 2. *This figure illustrates the mapping* $\theta(\cdot)$. *The (discontinuous) solid line is the mapping while the dash line represents the linear map induced by* G. *The dash-dot line represents the deviation band controlled by* r_{1n} *and* R_{2n}.

by rational decision-makers; see e.g. Hansen and Singleton [21], Chamberlain [10], Imbens [25], and Donald, Imbens and Newey [13]. Thus, the curved exponential framework is a fundamental complement of the exponential framework, at least in certain fields of data analysis.

We require the following additional regularity conditions on the mapping $\theta(\cdot)$:

NE.3 For every κ, and uniformly in $\gamma \in B(0, \kappa\sqrt{d})$, there exists a linear operator $G : \mathbb{R}^{d_1} \to \mathbb{R}^d$ such that $G'G$ has eigenvalues bounded from above and away from zero, uniformly in n, and for every n

$$\sqrt{n}\left(\theta(\eta_0 + \gamma/\sqrt{n}) - \theta(\eta_0)\right) = r_{1n} + (I_d + R_{2n})G\gamma,$$

where $\|r_{1n}\| \leq \delta_{1n}$ and $\|R_{2n}\| \leq \delta_{2n}$ and $\delta_{1n}\sqrt{d} \to 0$ and $\delta_{2n}d \to 0$.

Thus the mapping $\eta \mapsto \theta(\eta)$ is allowed to be nonlinear and discontinuous. For example, the additional condition of $\delta_{1n} = 0$ implies the continuity of the mapping in a neighborhood of η_0. More generally, condition NE.3 does impose that the map admits an approximate linearization in the neighborhood of η_0, whose quality is controlled by the errors δ_{1n} and δ_{2n}. An example of a kind of map allowed in this framework is given in Figure 2.

Given a prior π on Θ, the posterior of η given the data is denoted by

$$\pi_n(\eta) \propto \pi(\theta(\eta)) \cdot \prod_{i=1}^n h(X_i; \eta) = \pi(\theta(\eta)) \cdot \exp\left(n\bar{X}'\theta(\eta) - n\psi(\theta(\eta))\right).$$

In this framework, we also define the local parameters to describe contiguous deviations from the true parameter as

$$\gamma = \sqrt{n}(\eta - \eta_0) - s, \quad s = (G'G)^{-1}G'\sqrt{n}(\bar{x} - \mu),$$

where s is a first order approximation to the normalized maximum likelihood/extremum estimate. Further, we have that $E[s] = 0$, $E[ss'] = (G'G)^{-1}$, and $\|s\| = O_p(\sqrt{d})$.

The posterior density of γ over Γ, where $\Gamma = \sqrt{n}(\Psi - \eta_0) - s$, is $f(\gamma) = \frac{\ell(\gamma)}{\int_\Gamma \ell(\gamma)d\gamma}$, where

$$
\begin{aligned}
\ell(\gamma) = \ & \exp\left(n\bar{X}'\left(\theta\left(\eta_0 + \frac{\gamma+s}{\sqrt{n}}\right) - \theta\left(\eta_0 + \frac{s}{\sqrt{n}}\right)\right)\right) \\
& \times \exp\left(-n\psi\left(\theta\left(\eta_0 + \frac{\gamma+s}{\sqrt{n}}\right)\right) + n\psi\left(\theta\left(\eta_0 + \frac{s}{\sqrt{n}}\right)\right)\right) \\
& \times \pi\left(\theta\left(\eta_0 + \frac{\gamma+s}{\sqrt{n}}\right)\right) / \pi\left(\theta\left(\eta_0 + \frac{s}{\sqrt{n}}\right)\right).
\end{aligned}
\tag{5.31}
$$

The condition on the prior is the following:

NE.4 The prior $\pi(\eta) \propto \pi(\theta(\eta))$, where $\pi(\theta)$ satisfies condition E.3.

Theorem 5 *Conditions E.1-E.4 and NE.1-NE.4 imply conditions C.1-C.3 with* $\|K\| = C\sqrt{d/\lambda_{min}}$ *for some* $C > 1$, *where* $\lambda_{\min}$ *is the minimal eigenvalue of* $J = G'G$.

Comment 5.2 *Theorems 1 and 5 imply the asymptotic normality of the posterior,*

$$
\int_\Gamma |f(\gamma) - \phi(\gamma)|d\gamma = o_p(1),
$$

where

$$
\phi(\gamma) = \frac{1}{(2\pi)^{d/2} \det\left((G'G)^{-1}\right)^{1/2}} \exp\left(-\frac{1}{2}\gamma'(G'G)\gamma\right).
$$

Theorem 2 implies further that the main results of the paper on the polynomial time sampling and integration apply to this curved exponential family.

5.2. Z-estimation

Next we turn to the Z-estimation problem, where our basic setup closely follows the setup in e.g. He and Shao [22]. We make the following assumption that characterizes the setting. As in the rest of the paper, the dimension of the parameter space d and other quantities will depend on the sample size n.

ZE.0 The data $X_1, ..., X_n$ are i.i.d, and there exists a vector-valued moment function $m : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}^{d_1}$ such that

$$
E[m(X, \theta)] = 0 \ \text{ at the true parameter } \theta = \theta_0 \in \Theta_n \subset B(\theta_0, T_n) \subset \mathbb{R}^d.
$$

Both the dimension of the moment function d_1 and the dimension of the parameter d grow with the sample size n, and we restrict that $cd_1 \leq d \leq d_1$ for some constant c. The parameter space Θ_n is an open convex set contained in the ball $B(\theta_0, T_n)$ of radius T_n, where the radius T_n can grow with the sample size n.

The normalized empirical moment function takes the form

$$S_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m(X_i, \theta).$$

The Z-estimator for θ_0 is defined as the minimizer of the norm $\|S_n(\theta)\|$. However, in many applications of interests, the lack of continuity or smoothness of the empirical moments $S_n(\theta)$ can pose serious computational challenges to obtaining the minimizer. As argued in the introduction, in such cases the MCMC methodology could be particularly appealing for obtaining the quasi-posterior means and medians as computationally tractable alternatives to the Z-estimator based on minimization.

We then make the following variance and smoothness assumptions on the moment functions in addition to the basic condition **ZE.0**:

ZE.1 Let $S^{d_1} = \{\eta \in \mathbb{R}^{d_1} : \|\eta\| = 1\}$ denote the unit sphere. The variance of the moment function is bounded, namely $\sup_{\eta \in S^{d_1}} E[(\eta'm(X, \theta_0))^2] = O(1)$. The moment functions have the following continuity property: $\sup_{\eta \in S^{d_1}} (E[(\eta'(m(X, \theta) - m(X, \theta_0)))^2])^{1/2} \le O(1) \cdot \|\theta - \theta_0\|^{\alpha}$, uniformly in $\theta \in \Theta_n$, where $\alpha \in (0, 1]$ and is bounded away from zero, uniformly in n. Moreover, the family of functions $\mathcal{F} = \{\eta'(m(X, \theta) - m(X, \theta_0)) : \theta \in \Theta_n \subset \mathbb{R}^d, \eta \in S^{d_1}\}$ is not very complex, namely the uniform covering entropy of $\mathcal{F}$ is of the same order as the uniform covering entropy of a Vapnik-Chervonenkis (VC) class of functions with VC dimension of order $O(d)$, and $\mathcal{F}$ has an envelope F a.s. bounded by $M = O(\sqrt{d})$.

The smoothness assumption covers moment function both in the smooth case, where $\alpha = 1$, and the non-smooth case, where $\alpha < 1$. For example, in the classical mean regression problem, we have the smooth case $\alpha = 1$ and in the quantile regression problems mentioned in the introduction, we have a non-smooth case, with $\alpha = 1/2$. The condition on the function class $\mathcal{F}$ is standard in statistical estimation and, in particular, holds for $\mathcal{F}$ formed as VC classes or certain stable transformations of VC classes (see van der Vaart and Wellner [49]). We use the entropy in conjunction with the maximal inequalities similar to those developed in He and Shao [22]. The condition on the envelope is standard, but it can be replaced by an alternative condition on $\sup_{f \in \mathcal{F}} n^{-1} \sum_{i=1}^{n} f^4$, see e.g. He and Shao [22], which can weaken the assumptions on the envelope.

Next we make the following additional smoothness and identification assumptions uniformly in the sample size n.

ZE.2 The mapping $\theta \mapsto E[m(X, \theta)]$ is continuously twice differentiable with $\|\sup_{\eta \in S^{d_1}} \nabla_\theta^2 E[m(X, \theta)][\eta, \eta]\|$ bounded by $O(\sqrt{d})$ uniformly in θ, uniformly in n. The eigenvalues of $A'A$, where $A = \nabla E[m(X, \theta_0)]$ is the Jacobian matrix, are bounded above and away from zero uniformly in n. Finally, there exist positive numbers μ and δ such that uniformly in n, the following identification condition holds

$$\|E[m(X, \theta)]\| \ge (\sqrt{\mu}\|\theta - \theta_0\| \ \wedge \ \delta). \tag{5.32}$$

This condition requires the population moments $E[m(X,\theta)]$ to be approximately linear in the parameter θ near the true parameter value θ_0, and also insures identifiability of the true parameter value θ_0.

Finally, we impose the following restrictions on the parameter dimension d and the radius of the parameter space T_n.

ZE.3 The following condition holds: (a) $d^4 \log^2 n/n \to 0$, (b) $d^{2+\alpha} \log n/n^\alpha \to 0$, and (c) $dT_n^{2\alpha} \log n/n \to 0$.

These conditions are reasonable. Indeed, if we set $\alpha = 1$ and use radius $T_n = O(d \log n)$ for parameter space, then we require only that $d^4/n \to 0$, ignoring logs, which is only slightly stronger than the condition $d^3/n \to 0$ needed in the exponential family case. In the latter case, the information on higher order moments lead to the weaker requirement. Also, an important difference here is that we are using the flat prior in the Z-estimation framework, and this necessitates us to restrict the radius of parameter space by T_n. Note that even though the bounded radius $T_n = O(1)$ is already plausible for many applications, we can allow for the radius to grow, for example, $T_n = O(d \log n)$ when $\alpha = 1$.

In order to state the formal results concerning the quasi-posterior, let us define the quasi-posterior and related quantities. First, we define the criterion function as $Q_n(\theta) = -\|S_n(\theta)\|^2$, and treat it as a replacement for the log-likelihood. We will use a flat prior over the parameter space Θ, so that the quasi-posterior density of θ over Θ takes the form

$$\pi_n(\theta) = \frac{\exp\{Q_n(\theta)\}}{\int_\Theta \exp\{Q_n(\theta')\}d\theta'}.$$

We associate every point θ in the parameter space Θ with a local parameter $\lambda \in \Lambda = \sqrt{n}(\Theta - \theta_0) - s$, where $\lambda = \sqrt{n}(\theta - \theta_0) - s$, and $s = -(A'A)^{-1}A'S_n(\theta_0)$ is a first order approximation to extremum estimate. We have that $E[m(X,\theta_0)m(X,\theta_0)']$ is bounded in the spectral norm, and $(A'A)^{-1}A'$ has a bounded norm, so that the norm of s can be bounded in probability, $\|s\| = O_p(\sqrt{d})$, by the Chebyshev inequality. Finally, the quasi-posterior density of λ over $\Lambda = \sqrt{n}(\Theta - \theta_0) - s$ is given by

$$f(\lambda) = \ell(\lambda)/\int_\Lambda \ell(\lambda')d\lambda',$$

where

$$\ell(\lambda) = \exp(Q_n(\theta_0 + (\lambda + s)/\sqrt{n}) - Q_n(\theta_0 + s/\sqrt{n})).$$

Theorem 6 *Conditions ZE.0-ZE.3 imply conditions C.1-C.3 with $\|K\| = C\sqrt{d/\lambda_{min}}$ for $C > 1$, where $\lambda_{\min}$ is the minimal eigenvalue of $J = 2A'A$.*

Comment 5.3 *Theorems 1 and 6 imply the asymptotic normality of the quasi-posterior,*

$$\int_\Lambda |f(\lambda) - \phi(\lambda)|d\lambda = o_p(1),$$

where

$$\phi(\lambda) = \frac{1}{(2\pi)^{d/2} \det J^{1/2}} \exp\left(-\frac{1}{2}\lambda'J\lambda\right).$$

Theorem 2 implies further that the main results of the paper on the polynomial time sampling and integration apply to the quasi-posterior density formulated for the Z-estimation framework.

6. Conclusion

In this paper we study the implications of the statistical large sample theory for computational complexity of Bayesian and quasi-Bayesian estimation carried out using a canonical Metropolis random walk. Our analysis permits the parameter dimension of the problem to grow to infinity and allows the underlying log-likelihood or extremum criterion function to be discontinuous and/or non-concave. We establish polynomial complexity by exploiting a central limit theorem framework which provides the structural restriction on the problem, namely, that the posterior or quasi-posterior density approaches a normal density in large samples.

We focused the analysis on (general) Metropolis random walks and provided specific bounds for a canonical Gaussian random walk. Although it is widely used for its simplicity, this canonical random walk is not the most sophisticated algorithm available. Thus, in principle further improvements could be obtained by considering different kinds of algorithms, for example, the Langevin diffusion [43, 48, 45, 1]. (Of course, the algorithm requires a smooth gradient of the log-likelihood function, which rules out the nonsmooth and discontinuous cases emphasized here.) Another important research direction, as suggested by a referee, could be to develop sampling and integration algorithms that most effectively exploit the proximity of the posterior to the normal distribution.

Acknowledgement

Appendix A: Proofs of Other Results

Proof of Theorem 1. From C.1 it follows that

$$
\begin{aligned}
\int_\Lambda |f(\lambda) - \phi(\lambda)| d\lambda &\leq \int_K |f(\lambda) - \phi(\lambda)| d\lambda + \int_{K^c} (f(\lambda) + \phi(\lambda))\, d\lambda \\
&= \int_K |f(\lambda) - \phi(\lambda)| d\lambda + o_p(1)
\end{aligned}
$$

Now, denote $C_n = \dfrac{(2\pi)^{d/2} \det\left(J^{-1}\right)^{1/2}}{\int_K \ell(\omega)d\omega}$ and write

$$\int_K \left| \frac{f(\lambda)}{\phi(\lambda)} - 1 \right| \phi(\lambda)d\lambda \;=\; \int_K \left| C_n \cdot \exp\left(\ln \ell(\lambda) - \left(-\frac{1}{2}\lambda' J\lambda \right) \right) - 1 \right| \phi(\lambda)d\lambda$$

Combining the expansion in C.2 with conditions imposed in C.3,

$$
\begin{aligned}
\int_\Lambda \left| \frac{f(\lambda)}{\phi(\lambda)} - 1 \right| \phi(\lambda)d\lambda \;&\leq\; \int_K |C_n \cdot \exp\left(\epsilon_1 + \epsilon_2\lambda' J\lambda \right) - 1| \, \phi(\lambda)d\lambda \\
&\quad + \int_K |C_n \cdot \exp\left(-\epsilon_1 - \epsilon_2\lambda' J\lambda \right) - 1| \, \phi(\lambda)d\lambda \\
&\leq\; 2\int_K \left| C_n \cdot e^{o_p(1)} - 1 \right| \phi(\lambda)d\lambda \\
&\leq\; 2|C_n e^{o_p(1)} - 1|
\end{aligned}
$$

The proof then follows by showing that $C_n \to_p 1$. Using condition C.1 on the set $K = B(0, \|K\|)$ and C.2,

$$
\begin{aligned}
\frac{1}{C_n} \;&\geq\; \frac{\displaystyle\int_K \ell(\lambda)d\lambda}{(1 + o(1)) \displaystyle\int_K e^{-\frac{1}{2}\lambda' J\lambda}d\lambda} \;\geq\; \frac{\displaystyle\int_K e^{-\frac{1}{2}\lambda' J\lambda} e^{-\epsilon_1 - \frac{\epsilon_2}{2}(\lambda' J\lambda)}d\lambda}{(1 + o(1)) \displaystyle\int_K e^{-\frac{1}{2}\lambda' J\lambda}d\lambda} \\[2mm]
&=\; \frac{e^{-\epsilon_1}}{(1 + o(1))} \sqrt{\frac{\det(J)}{\det(J + \epsilon_2 J)}} \frac{\displaystyle\int_K \frac{e^{-\frac{1}{2}\lambda'(J + \epsilon_2 J)\lambda}}{(2\pi)^{d/2} \det((J + \epsilon_2 J)^{-1})^{1/2}}d\lambda}{\displaystyle\int_K \frac{e^{-\frac{1}{2}\lambda' J\lambda}}{(2\pi)^{d/2} \det(J^{-1})^{1/2}}d\lambda}.
\end{aligned}
$$

Since $\epsilon_2 < 1/2$, we can define $W \sim N(0, (1+\epsilon_2)^{-1}J^{-1})$ and $V \sim N(0, J^{-1})$ and rewrite our bound as

$$
\begin{aligned}
\frac{1}{C_n} \;&\geq\; \frac{e^{-\epsilon_1}}{(1 + o(1))} \left(\frac{1}{1 + \epsilon_2} \right)^{d/2} \frac{P(\|W\| \leq \|K\|)}{P(\|V\| \leq \|K\|)} \\
&\geq\; \frac{e^{-\epsilon_1}}{(1 + o(1))} \left(\frac{1}{1 + \epsilon_2} \right)^{d/2}
\end{aligned}
$$

where the last inequality follows from $P(\|W\| \leq \|K\|) \geq P(\|\sqrt{1 + \epsilon_2}W\| \leq \|K\|) = P(\|V\| \leq \|K\|)$. Likewise,

$$\frac{1}{C_n} \leq \frac{\displaystyle\int_K \ell(\lambda)d\lambda}{\displaystyle\int_K e^{-\frac{1}{2}\lambda' J\lambda}d\lambda} \;\leq\; e^{\epsilon_1} \left(\frac{1}{1 - \epsilon_2} \right)^{d/2}$$

Therefore $C_n \to_p 1$ since $\epsilon_1 \to_p 0$ and $\epsilon_2 \cdot d \to_p 0$ (cf. Comment 2.1). ■

Proof of Lemma 1. The result follows immediately from equations (2.8)-(2.9). ■

Proof of Lemma 2. Let $M := \beta \frac{2te^{-t^2/4}}{\sqrt{\pi}}$. Take any measurable partition of $K = S_1 \cup S_2 \cup S_3$, with $d(S_1, S_2) \geq t$. It suffices to prove that

$$\int \left(M 1_{S_i}(x) - 1_{S_3}(x) \right) f(x) dx < 0, \text{ for } i = 1 \text{ or } i = 2.$$

We will prove this by contradiction. Suppose that

$$\int \left(M 1_{S_i}(x) - 1_{S_3}(x) \right) f(x) dx > 0, \text{ for } i = 1 \text{ and } i = 2.$$

We will use the Localization Lemma of Kannan, Lovász, and Simonovits [29] in order to reduce a high-dimensional integral to a low-dimensional integral.

Lemma 6 (Localization Lemma) *Let g and h be two lower semi-continuous Lebesgue integrable functions on $\mathbb{R}^d$ such that*

$$\int_{\mathbb{R}^d} g(x) dx > 0 \quad and \quad \int_{\mathbb{R}^d} h(x) dx > 0.$$

Then there exist two points $a, b \in \mathbb{R}^d$, and a linear function $\tilde{\gamma} : [0,1] \to \mathbb{R}_+$ such that

$$\int_0^1 \tilde{\gamma}^{d-1}(t) g((1-t)a + tb) dt > 0 \quad and \quad \int_0^1 \tilde{\gamma}^{d-1}(t) h((1-t)a + tb) dt > 0,$$

where $([a, b], \tilde{\gamma})$ is said to form a needle.

Proof. See Kannan, Lovász, and Simonovits [29]. ∎

By the Localization Lemma, there exists a needle $(a, b, \tilde{\gamma})$ such that

$$\int_0^1 \tilde{\gamma}^{d-1}(l) f((1-l)a + lb) \left(M 1_{S_i}((1-l)a + lb) - 1_{S_3}((1-l)a + lb) \right) du > 0,$$

for $i = 1, 2$. Equivalently, using $\gamma(u) = \tilde{\gamma}(u/\|b - a\|)$ and $v := (b - a)/\|b - a\|$ where $\|b - a\| \geq t$, and rearranging we have for $i = 1, 2$,

$$M \int_0^{\|b-a\|} \gamma^{d-1}(u) f(a + uv) 1_{S_i}(a + uv) du$$
$$> \int_0^{\|b-a\|} \gamma^{d-1}(u) 1_{S_3}(a + uv) f(a + uv) du. \tag{A.33}$$

In order for the left hand side of (A.33) be positive for $i = 1$ and $i = 2$, the line segment $[a, b]$ must contain points in S_1 and S_2. Since $d(S_1, S_2) \geq t$, we have that $S_3 \cap [a, b]$ contains an interval $[w, w + t]$ whose length is at least t. Thus, we can partition the line segment $[a, b]$ into $[0, w) \cup [w, w + t] \cup (w + t, \|b - a\|]$. We will prove that for every $w \in \mathbb{R}$ such that $0 \leq w \leq w + t \leq \|b - a\|$

$$\int_w^{w+t} \gamma^{d-1}(u) f(a + uv) du \geq M \min \left\{ \int_0^w \gamma^{d-1}(u) f(a + uv) du, \right.$$
$$\left. \int_{w+t}^{\|b-a\|} \gamma^{d-1}(u) f(a + uv) du \right\} \tag{A.34}$$

which contradicts the relation (A.33) and proves the lemma.

First, note that $f(a + uv) = e^{-\|a+uv\|^2} m(a + uv) = e^{-u^2 + r_1 u + r_0} m(a + uv)$ where $r_1 := 2a'v$ and $r_0 := -\|a\|^2$. Next, recall that $m(a + uv)\gamma^{d-1}(u)$ is still a unidimensional log-β-concave function on u. By Lemma 9 presented in Appendix B, there exists a unidimensional logconcave function $\widehat{m}$ such that $\beta\widehat{m}(u) \leq m(a+uv)\gamma^{d-1}(u) \leq \widehat{m}(u)$ for every u. Moreover, there exists numbers s_0 and s_1 such that $\widehat{m}(w) = s_0 e^{s_1 w}$ and $\widehat{m}(w + t) = s_0 e^{s_1(w+t)}$. Due to the log-concavity of $\widehat{m}$, this implies that

$$\widehat{m}(u) \geq s_0 e^{s_1 u} \ \text{ for } u \in (w, w+t) \ \text{ and } \ \widehat{m}(u) \leq s_0 e^{s_1 u} \ \text{ otherwise.}$$

Thus, if we replace $m(a + uv)\gamma^{d-1}(u)$ by $s_0 e^{s_1 u}$ on the right hand side of (A.34) and replace $m(a+uv)\gamma^{d-1}(u)$ by $\beta s_0 e^{s_1 u}$ on the left hand side of (A.34), and define $\widehat{r}_1 = r_1 + s_1$ and $\widehat{r}_0 := r_0 + \ln s_0$, we obtain the relation

$$\beta \int_w^{w+t} e^{-u^2 + \widehat{r}_1 u + \widehat{r}_0} du \geq M \min \left\{ \int_0^w e^{-u^2 + \widehat{r}_1 u + \widehat{r}_0} du, \ \int_{w+t}^{\|b-a\|} e^{-u^2 + \widehat{r}_1 u + \widehat{r}_0} du \right\}.$$

This relation is stronger than (A.34) and thus implies (A.34). This relation is equivalent to

$$\beta \int_w^{w+t} e^{-(u-\frac{\widehat{r}_1}{2})^2 + \widehat{r}_0 + \frac{\widehat{r}_1^2}{4}} du \quad \geq M \min \left\{ \int_0^w e^{-(u-\frac{\widehat{r}_1}{2})^2 + \widehat{r}_0 + \frac{\widehat{r}_1^2}{4}} du, \right.$$
$$\left. \int_{w+t}^{\|b-a\|} e^{-(u-\frac{\widehat{r}_1}{2})^2 + \widehat{r}_0 + \frac{\widehat{r}_1^2}{4}} du \right\}. \tag{A.35}$$

Now, cancel the term $e^{\widehat{r}_0 + \widehat{r}_1^2/4}$ on both sides and, since we want the inequality (A.35) holding for any w, (A.35) is implied by

$$\int_w^{w+t} e^{-u^2} du \geq \frac{2t e^{-t^2/4}}{\sqrt{\pi}} \min \left\{ \int_{-\infty}^w e^{-u^2} du, \ \int_{w+t}^{\infty} e^{-u^2} du \right\} \tag{A.36}$$

holding for any w. This inequality is Lemma 2.2 in Kannan and Li [28].∎

Proof of Corollary 1. Consider the change of variables $\tilde{x} = \frac{J^{1/2}x}{\sqrt{2}}$ and $\tilde{S} = \frac{J^{1/2}S}{\sqrt{2}}$. Then, in $\tilde{x}$ coordinates, $f(\tilde{x}) = e^{\tilde{x}'\tilde{x}} m(\sqrt{2}J^{-1/2}\tilde{x})$ satisfies the assumption of Lemma 2 and $d(\tilde{S}_1, \tilde{S}_2) \geq \frac{t\sqrt{\lambda_{min}}}{\sqrt{2}}$. The result follows by applying Lemma 2 with $\tilde{x}$ coordinates. ∎

Proof of Lemma 3. The result is immediate from the stated assumptions. ∎

Proof of Theorem 2. See section 3.2. ∎

Proof of Lemma 4. Define $K := B(0, R)$, so that R is the radius of K; also let $r := 4\sqrt{d}\sigma$ (where $\sigma^2 \leq \frac{1}{16dL^2}$), and let $q(x|u)$ denote the normal density function centered at u with covariance matrix $\sigma^2 I$. We use the following notation: $B_u = B(u, r)$, $B_v = B(v, r)$, and $A_{u,v} = B_u \cap B_v \cap K$. By definition of r, we have that $\int_{B_u} q(x|u)dx = \int_{B_v} q(x|v)dx \geq 1 - P\{|U| \geq 4\} > 1 - 1/10^4$, where $U \sim N(0, 1)$.

Define the direction $w = (v-u)/\|v-u\|$. Let $H_1 = \{x \in B_u \cap B_v : w'(x-u) \geq \|v - u\|/2\}$, $H_2 = \{x \in B_u \cap B_v : w'(x - u) \leq \|v - u\|/2\}$. Consider the one-step distributions from u and v. We first observe that in view of Lemma 1 and Lemma 3 that $\inf_{x \in B(y,r)} f(x)/f(y) \geq \beta e^{-Lr}$. Then we have that

$$\|P_u - P_v\|_{TV} \leq 1 - \int_K \min\{dP_u, dP_v\} \leq 1 - \int_{A_{u,v}} \min\{dP_u, dP_v\}$$

$$= 1 - \int_{A_{u,v}} \min\left\{ q(x|u) \min\left\{ \frac{f(x)}{f(u)}, 1 \right\}, q(x|v) \min\left\{ \frac{f(x)}{f(v)}, 1 \right\} \right\} dx$$

$$\leq 1 - \beta e^{-Lr} \int_{A_{u,v}} \min\left\{ q(x|u), q(x|v) \right\} dx$$

$$\leq 1 - \beta e^{-Lr} \left(\int_{H_1 \cap K} q(x|u)dx + \int_{H_2 \cap K} q(x|v)dx \right),$$

where $\|u - v\| < \sigma/8$. Next we will bound from below the last sum of integrals for an arbitrary $u \in K$.

We first bound the integrals over the possibly larger sets, respectively H_1 and H_2. Let h denote the density function of a univariate random variable distributed as $N(0, \sigma^2)$. It is easy to see that $h(t) = \int_{w'(x-u)=t} q(x|u)dx$, i.e. h is the marginal density of $q(\cdot|u)$ along the direction w up to a translation. Let $H_3 = \{x : -\|u - v\|/2 < w'(x - u) < \|v - u\|/2\}$. Note that $B_u \subset H_1 \cup (H_2 - \|u - v\|w) \cup H_3$ where the union is disjoint. Armed with these observations, we have

$$
\begin{aligned}
\int_{H_1} q(x|u)dx + \int_{H_2} q(x|v)dx &= \int_{H_1} q(x|u)dx + \int_{H_2 - \|u-v\|w} q(x|u)dx \\
&\geq \int_{B_u} q(x|u)dx - \int_{H_3} q(x|u)dx \\
&= \int_{B_u} q(x|u)dx - \int_{-\|u-v\|/2}^{\|u-v\|/2} h(t)dt \\
&\geq 1 - \frac{1}{10^4} - \int_{-\|u-v\|/2}^{\|u-v\|/2} \frac{e^{-t^2/2\sigma^2}}{\sqrt{2\pi}\sigma} dt \\
&\geq 1 - \frac{1}{10^4} - \|u - v\| \frac{1}{\sqrt{2\pi}\sigma} \\
&\geq 1 - \frac{1}{10^4} - \frac{1}{8\sqrt{2\pi}} \geq \frac{9}{10}, \qquad (A.37)
\end{aligned}
$$

where we used that $\|u - v\| < \sigma/8$ by the hypothesis of the lemma.

In order to take the support K into account, we can assume that $u, v \in \partial K$, i.e. $\|u\| = \|v\| = R$ (otherwise the integral will be larger). Let $z = (v + u)/2$ and define the half space $H_z = \{x : z'x \leq z'z\}$ whose boundary passes through u and v (Using $\|u\| = \|v\| = R$ it follows that $z'v = z'u = z'z/2$).

By the symmetry of the normal density, we have

$$\int_{H_1 \cap H_z} q(x|u)dx = \frac{1}{2} \int_{H_1} q(x|u)dx.$$

Although $H_1 \cap H_z$ does not lie in K in general, simple arithmetic shows that $H_1 \cap \left(H_z - \frac{r^2 z}{R\|z\|} \right) \subseteq K$.[1]

Using that $\int_{H_z \backslash (H_z - \frac{r^2 z}{R\|z\|})} q(x|u) = \int_0^{r^2/R} h(t)dt$, we have

$$
\begin{aligned}
\int_{H_1 \cap K} q(x|u)dx \;\; & \geq \;\; \int_{H_1 \cap \left(H_z - \frac{r^2 z}{R\|z\|} \right)} q(x|u)dx \geq \int_{H_1 \cap H_z} q(x|u)dx - \int_0^{r^2/R} h(t)dt \\
& \geq \;\; \frac{1}{2} \int_{H_1} q(x|u)dx - \int_0^{r^2/R} \frac{e^{-t^2/2\sigma^2}}{\sqrt{2\pi}\sigma} dt \\
& \geq \;\; \frac{1}{2} \int_{H_1} q(x|u)dx - 4\sqrt{d}\sigma \frac{1}{30\sqrt{d}} \frac{1}{\sqrt{2\pi}\sigma},
\end{aligned}
$$

where we used that $\frac{r}{R} < \frac{1}{30\sqrt{d}}$ since $r = 4\sqrt{d}\sigma$ and $\frac{\sigma}{R} < \frac{1}{120d}$.

By symmetry, the same inequality holds when u and H_1 are replaced by v and H_2 respectively. Adding these inequalities and using (A.37), we have

$$
\left(\int_{H_1 \cap K} q(x|u)dx + \int_{H_2 \cap K} q(x|v)dx \right) \geq \frac{9}{20} - \frac{4}{15\sqrt{2\pi}} \geq 1/3. \qquad \text{(A.38)}
$$

Thus, we have

$$
\|P_u - P_v\| < 1 - \frac{\beta}{3} e^{-Lr}
$$

and the result follows since $Lr \leq 1$. $\blacksquare$

Proof of Lemma 5. We calculate the probability p of making a proper move. We will use the notation defined in the proof of Lemma 4. Let u be an arbitrary point in K. We have that

$$
p_u \;\; = \;\; \int_K \min \left\{ \frac{f(x)}{f(u)}, \; 1 \right\} q(x|u)dx \geq \beta e^{-Lr} \int_{B_u \cap K} q(x|u)dx \geq \beta e^{-Lr} \frac{1}{3},
$$

where we used that $\inf_{x \in B(y,r)} f(x)/f(y) \geq \beta e^{-Lr}$ by Lemma 1 and Lemma 3 and the bound (A.38) for the case that $u = v$ so that $B_u = H_1 \cup H_2$. Since $Lr < 1$ we conclude that $p_u \geq \beta/3e$.

We then note that for $Q(A) > 0$ the ratio $Q_0(A)/Q(A)$ is bounded above by $\sup_{x \in K} dQ_0(x)/dQ(x)$; $dQ_0(x)/dx$ is bounded above by $p_u^{-1} e^{-\|x\|^2/2\sigma^2} \cdot (2\pi\sigma^2)^{-d/2} \leq p_u^{-1} \cdot (2\pi\sigma^2)^{-d/2}$; and $dQ(x)/dx$ is bounded over $x \in K$ below by $(2\pi)^{-d/2} \det(J^{1/2}) e^{-\frac{1}{2}x'Jx} \beta^{1/2} \geq (2\pi)^{-d/2} \lambda_{min}^{d/2} e^{-\frac{1}{2}\|K\|_J^2} \beta^{1/2}$, where $\beta = e^{-2(\epsilon_1 + \epsilon_2\|K\|_J^2/2)}$.

---

[1] Indeed, take $y \in H_1 \cap \left(H_z - \frac{r^2}{R} \frac{z}{\|z\|} \right)$. We can write $y = \frac{z}{\|z\|} \left(\frac{y'z}{\|z\|} \right) + s$, where $\|s\| \leq r$ (since $\left\| y - \frac{z}{\|z\|} \left(\frac{y'z}{\|z\|} \right) \right\| \leq \|y - z\| = \|y - \frac{u+v}{2}\| \leq \frac{1}{2}\|y - u\| + \frac{1}{2}\|y - v\| \leq r$) and s is also orthogonal to z. Since $y \in \left(H_z - \frac{r^2}{R} \frac{z}{\|z\|} \right)$, we have $\frac{y'z}{\|z\|} \leq \frac{z'z}{\|z\|} - \frac{r^2}{R} = \|z\| - \frac{r^2}{R} \leq R - \frac{r^2}{R}$. Therefore, $\|y\| = \sqrt{\left(\frac{y'z}{\|z\|} \right)^2 + \|s\|^2} \leq \sqrt{(R - \frac{r^2}{R})^2 + r^2} = \sqrt{R^2 - r^2(1 - \frac{r^2}{R^2})} \leq R$.

Thus, we can bound

$$
\begin{aligned}
\max_{A\in\mathcal{A}:Q(A)>0}\tfrac{Q_0(A)}{Q(A)} \quad &\leq p_u^{-1}\sigma^{-d}\lambda_{min}^{-d/2}e^{\frac{1}{2}\|K\|_J^2}\beta^{-1/2}\\
&\leq 3e[120\sqrt{d}\lambda_{max}\|K\|/\sqrt{\lambda_{min}}]^d e^{\frac{1}{2}\|K\|_J^2}\beta^{-3/2}\\
&\leq 3[120\|K\|_J^2]^d e^{3\epsilon_1+2\epsilon_2\|K\|_J^2+1},
\end{aligned}
$$

where we used the bound on σ given in (3.22), and the fact that $\|K\|_J \geq \sqrt{\lambda_{\min}}\|K\|$ and $\|K\|_J > \sqrt{d}\,\sqrt{\lambda_{max}/\lambda_{min}}$ (cf. Comment 2.1).

The remaining results in the Lemma follow by invoking the CLT conditions.

■

Proof of Theorem 3. We have that, for λ^B denoting the random variable with law Q_B and λ denoting the random variable with law Q, and $MSE(\widehat{\mu}_g|X)$ denoting the mean square error $E[(\widehat{\mu}_g-\mu_g)^2|X]$ conditional on the element $\lambda^{0,B}$ drawn according to $X = \lambda^B$ or $X = \lambda$:

$$
\begin{aligned}
MSE(\widehat{\mu}_g) \quad &= \quad E_{Q_B}\left[MSE(\widehat{\mu}_g|\lambda^B)\right] = E_Q\left[MSE(\widehat{\mu}_g|\lambda)\frac{dQ_B(\lambda)}{dQ(\lambda)}\right]\\
&= \quad E_Q\left[MSE(\widehat{\mu}_g|\lambda)\right] + E_Q\left[MSE(\widehat{\mu}_g|\lambda)\left(\frac{dQ_B(\lambda)}{dQ(\lambda)}-1\right)\right]\\
&\leq \quad E_Q\left[MSE(\widehat{\mu}_g|\lambda)\right] + 4\bar{g}^2 E_Q\left[\left|\frac{dQ_B(\lambda)}{dQ(\lambda)}-1\right|\right]\\
&= \quad (\sigma_{g,N}^2/N) + 8\bar{g}^2\|Q_B - Q\|_{TV},
\end{aligned}
$$

where $\sigma_{g,N}^2$ is N times the variance of the sample average when the Markov chain starts from the stationary distribution Q. We also used the fact that $\|Q_B - Q\|_{TV} = \frac{1}{2}\int |dQ_B/dx - dQ/dx|dx$.

The bound on $\sigma_{g,N}^2$ will depend on the particular scheme, as discussed below. We begin by bounding the burn-in period B.

We require that the second term in the bound for $MSE(\widehat{\mu}_g)$ to be smaller than $\varepsilon/3$, which is equivalent to imposing that $\|Q_B - Q\|_{TV} < \frac{\varepsilon}{24\bar{g}^2}$. Using the conductance theorem of [37] restated in equation (3.15), since Q_0 is M-warm with respect to Q, we require that

$$
\sqrt{M}\left(1-\tfrac{\phi^2}{2}\right)^B \quad \leq \quad \sqrt{M}e^{-B\frac{\phi^2}{2}} \leq \frac{\varepsilon}{24\bar{g}^2} \text{ or } B \geq \frac{2}{\phi^2}\ln\left(\frac{24\sqrt{M}\bar{g}^2}{\varepsilon}\right).
$$

Next we bound $\sigma_{g,N}^2$. Specifically, we determine the number of post-burn-in iterations N_{lr}, N_{ss}, or N_{ms} needed to set $MSE(\widehat{\mu}_g) \leq \varepsilon$.

1. To bound N_{lr}, note that $\sigma_{g,N}^2 \leq \gamma_0\frac{4}{\phi^2}$ where the last inequality follows from the conductance-based covariance bound of [37] restated in equation (4.29). Thus, $N_{lr} = \frac{\gamma_0}{\varepsilon}\frac{6}{\phi^2}$ and B set above suffice to obtain $MSE(\widehat{\mu}_g) \leq \varepsilon$.

2. To bound N_{ss}, we first must choose a spacing S to ensure that the auto-covariances are sufficiently small. We start by bounding

$$
\sigma_{g,N}^2 \leq \gamma_0 + 2N|\gamma_S| \leq \gamma_0 + 2N\gamma_0\left(1-\frac{\phi^2}{2}\right)^S,
$$

where we used the conductance-based covariance bound of [37] restated in equation (4.29) and that $\lambda^{i,B}$ and $\lambda^{i+1,B}$ are spaced by S steps of the chain. By choosing S as

$$\left(1 - \frac{\phi^2}{2}\right)^S \le e^{-S\frac{\phi^2}{2}} \le \frac{\varepsilon}{6\gamma_0}, \text{ or } S \ge \frac{2}{\phi^2} \ln\left(\frac{6\gamma_0}{\varepsilon}\right),$$

and using $N_{ss} = \dfrac{3\gamma_0}{\varepsilon}$, we obtain

$$\begin{aligned} MSE(\widehat{\mu}_g) &\le \frac{1}{N_{ss}}\left(\gamma_0 + 2N_{ss}|\gamma_S|\right) + 8\bar{g}^2\|Q_B - Q\|_{TV} \\ &\le \frac{\varepsilon}{3\gamma_0}\left(\gamma_0 + 2\frac{3\gamma_0}{\varepsilon}\gamma_0\frac{\varepsilon}{6\gamma_0}\right) + \bar{g}^2\frac{\varepsilon}{3\bar{g}^2} \le \varepsilon \end{aligned}$$

3. To bound N_{ms}, we observe, using that $\lambda^{i,B}, i = 1, 2, ...,$ are i.i.d. across i, that $MSE(\widehat{\mu}_g) \le \frac{\gamma_0}{N_{ms}} + \varepsilon/3 \le \varepsilon$ provided that $N_{ms} \ge 2\gamma_0/(3\varepsilon)$. ∎

Proof of Theorem 4. Given

$$K = B(0, \|K\|) \text{ where } \|K\|^2 = cd,$$

condition C.1 holds by an argument given in proof of Ghosal's Lemma 4. Let $\lambda_n(c) = \sqrt{\frac{cd}{n}}B_{1n}(0) + \frac{cd}{n}B_{2n}(c)$. Our condition C.2 is satisfied by an argument similar to that given in the proof of Ghosal's Lemma 1 with

$$\epsilon_1 = O\left(\lambda_n(c)\|s\|^2\right) = O_p(\lambda_n(c)d) = O_p(d^{3/2}/n^{1/2}) = o_p(1) \text{ and}$$

$$\epsilon_2 = O\left(\lambda_n(c)\right) = O_p\left(d^{1/2}/n^{1/2}\right) = o_p(1/d),$$

and our condition C.3 is satisfied since $\epsilon_2\|K\|_J^2 = o_p(1)$. ∎

Comment A.1 *Ghosal [20] proves his results for the support set $K' = B(0, C\sqrt{d}\log d)$. His arguments actually go through for the support set $K = B(0, C\sqrt{d})$ due to the concentration of normal measure under $d \to \infty$ asymptotics. For details, see [4].*

Proof of Theorem 5. Take $K = B(0, \|K\|)$, where $\|K\|^2 = Cd_1$ for some C sufficiently large independent of d (see [4] for details). Let $\lambda_n(c) = \sqrt{\frac{cd}{n}}B_{1n}(0) + \frac{cd}{n}B_{2n}(c)$. Then condition C.1 is satisfied by the argument given in the proof of Ghosal's Lemma 4 and NE.3. Further, condition C.2 is satisfied by the argument similar to that given in the proof of Ghosal's Lemma 1 and by NE.3 with

$$\epsilon_1 = O_p\left(\delta_{1n}d^{1/2} + \delta_{2n}d + \lambda_n(C)(\delta_{1n}d^{1/2} + \delta_{2n}d^{1/2} + d)\right) = o_p(1),$$

$$\epsilon_2 = O_p\left(\lambda_n(C)\right) = o_p(d^{1/2}/n^{1/2}) = o_p(1/d),$$

and condition C.3 is satisfied since $\epsilon_2\|K\|_J^2 = o_p(1)$. ∎

Comment A.2 *For further details, see [4].*

Proof of Theorem 6. We will first establish the following linear approximation for $S_n(\theta)$ in a neighborhood of θ_0

$$\sup_{\|\theta-\theta_0\|\leq C\sqrt{d/n}} \|S_n(\theta) - S_n(\theta_0) - n^{1/2}A(\theta-\theta_0)\| = o_p\left(d^{-1/2}\right) \qquad (A.39)$$

for any fixed constant $C > 0$. For notational convenience let

$$\delta_n(\theta) = S_n(\theta) - S_n(\theta_0) - n^{1/2}A(\theta-\theta_0), \ W_n(\theta) = S_n(\theta) - S_n(\theta_0) - E\left[S_n(\theta) - S_n(\theta_0)\right]. \tag{A.40}$$

Let $\mathcal{F}_n = \{\eta'(m(X,\theta) - m(X,\theta_0)) \ : \|\theta - \theta_0\| \leq \rho_n, \eta \in S^{d_1}\}$. Under condition ZE.1, we apply the following maximal inequality adopted from He and Shao [22] (see [5] for details) to an empirical process indexed by members of $\mathcal{F}_n$:

$$\sup_{f\in\mathcal{F}_n} |n^{-1/2}\sum_{i=1}^{n} f(X_i)| = O_p\left(\sqrt{V}\sqrt{\log n}\left(\sup_{f\in\mathcal{F}_n} E[f^2] + n^{-1}VM^2\log n\right)^{1/2}\right). \tag{A.41}$$

Here the multiplier $\sqrt{V}$ arises as the order of the uniform bracketing entropy integral, where V is the VC dimension of a VC function class $\mathcal{F}_n$ or an entropically equivalent class $\mathcal{F}_n$. We assumed in ZE.1 that $V = O(d)$. Also M is the a.s. bound on the envelope of $\mathcal{F}_n$, assumed to be of order $O(\sqrt{d})$. Finally, we assumed that $\sup_{f\in\mathcal{F}_n}(E[f^2])^{1/2} = O(\rho_n^\alpha)$. Therefore, we have that uniformly in $\theta \in \Theta_n$

$$\begin{aligned} \|W_n(\theta)\| &= O_p\left(\sqrt{d\log n}\left(\|\theta-\theta_0\|^{2\alpha} + n^{-1}dM^2\log n\right)^{1/2}\right) \\ &= O_p\left(\sqrt{d\log n}\|\theta-\theta_0\|^\alpha + n^{-1/2}d^{3/2}\log n\right). \end{aligned} \tag{A.42}$$

Note that (A.42) and an expansion with an integral reminder around $\theta - \theta_0$ shows that uniformly in $\theta \in \Theta_n$

$$\begin{aligned} \|\delta_n(\theta)\| &\leq \|W_n(\theta)\| + \|\nabla^2 E[S_n(\xi)]\cdot[\theta-\theta_0,\theta-\theta_0]\| \\ &= O_p\left(d^{1/2}\log^{1/2}n\|\theta-\theta_0\|^\alpha + n^{-1/2}d^{3/2}\log n\right) + \\ &\quad + O_p\left(\sqrt{dn}\|\theta-\theta_0\|^2\right) \end{aligned}$$

where ξ lies between θ and θ_0 and we used ZE.2 that imposes $\|\nabla^2 E[S_n(\xi)]\cdot[\gamma,\gamma]\| = O(\sqrt{dn}\|\gamma\|^2)$. The condition (A.39) follows from the growth condition ZE.3(a).

Building upon (A.39), Lemmas 7 and 8 verify that conditions C.1-C.3 hold proving Theorem 6. ∎

Lemma 7 *Under conditions ZE.1-ZE.3, conditions C.2 and C.3 hold for $K = B(0, C\sqrt{d})$ for any fixed constant $C > 0$.*

Proof of Lemma 7. Let $s = -(A'A)^{-1}A'S_n(\theta_0)$ be a first order approximation for the extremum estimator. For $\theta = \theta_0 + (s+\lambda)/\sqrt{n}$ and $\tilde{\theta} = \theta_0 + s/\sqrt{n}$

$$
\begin{aligned}
\ln \ell(\lambda) &= -\|S_n(\theta)\|^2 + \|S_n(\tilde{\theta})\|^2 \\
&= -\lambda'A'A\lambda - \|r_n\|^2 - 2r_n'A\lambda - 2r_n'S_n(\tilde{\theta}) \\
&= -\lambda'A'A\lambda + o_p(1),
\end{aligned}
$$

where $r_n = \delta_n(\theta) - \delta_n(\tilde{\theta})$ for $\delta_n(\theta)$ defined in (A.40). Indeed, using (A.39) we have $\|\delta_n(\theta)\| = o_p(d^{-1/2})$ and $\|\delta_n(\tilde{\theta})\| = o_p(d^{-1/2})$ uniformly over $\lambda \in K$; using (A.39) we have $\|S_n(\tilde{\theta})\| = O_p(d^{1/2})$; and moreover, $\|\lambda\| = O(d^{1/2})$, and $\|s\| = O_p(d^{1/2})$ by Chebyshev inequality. Thus, conditions C.2 and C.3 follow with $\epsilon_1 = o_p(1)$, $\epsilon_2 = 0$, and $J = 2A'A$. ∎

Lemma 8 *Under the conditions ZE.1, ZE.2, and ZE.3 there exist a constant $C > 0$ such that by setting $K = B(0, C\sqrt{d})$ we have $\int_{K^c} \ell(\lambda)d\lambda = o_p\left(\int_K \ell(\lambda)d\lambda\right)$ and condition C.1 holds.*

Proof of Lemma 8. For notational convenience we conduct the proof in the original parameter space. Let $\tilde{\theta} = \theta_0 + s/\sqrt{n}$ and $\varepsilon > 0$ be any small positive constant. Since $\|s\| = O_p(d^{1/2})$, there is a constant $\widehat{C}$ such that $\|s\| \leq \widehat{C}d^{1/2}$, with asymptotic probability no smaller than $1 - \varepsilon$. Below we replace the last phrase by "wp $1 - \varepsilon$".

Now, since $E[S_n(\theta_0)] = 0$, we have that

$$
S_n(\theta) = W_n(\theta) + S_n(\theta_0) + E[S_n(\theta)], \tag{A.43}
$$

where $W_n(\theta)$ is defined in (A.42).

Next, define for $C \geq \widehat{C} + \tilde{C}$ the sets

$$
\widetilde{K} = B\left(\theta_0, \ \tilde{C}\sqrt{d/n}\right) \subseteq \widehat{K} = B\left(\tilde{\theta}, C\sqrt{d/n}\right), \tag{A.44}
$$

where the inclusion holds wp $1-\varepsilon$. Note that these sets are centered on different points. We will show that for a sufficiently large constant $\tilde{C}$

$$
\int_{\widehat{K}^c} \exp(-\|S_n(\theta)\|^2)d\theta = o_p\left(\int_{\widehat{K}} \exp(-\|S_n(\theta)\|^2)d\theta\right),
$$

which implies the claim of the lemma.

Step 1. Relative bound on $\|S_n(\theta_0)\|$. Note that $\|S_n(\theta_0)\| = O_p(d^{1/2})$ by Chebyshev inequality. Using equation (5.32) of condition ZE.2, we have that

$$
\|E[S_n(\theta)]\|^2 \geq \left(\sqrt{n}(\sqrt{\mu}\|\theta - \theta_0\| \wedge \delta)\right)^2 \geq \left(\tilde{C}\sqrt{\mu}\sqrt{d}\right)^2, \quad \forall \theta \in \tilde{K}^c
$$

since $\|\theta - \theta_0\| \geq \tilde{C}\sqrt{d/n}$. Therefore, there exists $\tilde{C}$ such that wp $1 - \varepsilon$

$$
\|E[S_n(\theta)]\| > 5\|S_n(\theta_0)\| \text{ uniformly in } \theta \in \widetilde{K}^c. \tag{A.45}
$$

Step 2. Relative bound on $\|W_n(\theta)\|$. Using equation (A.42), we have that for uniformly in $\theta \in \Theta_n \subset B(0, T_n)$

$$\|W_n(\theta)\| = O_p\left(\sqrt{d \log n}\|\theta - \theta_0\|^\alpha + n^{-1/2}d^{3/2}\log n\right),$$

Building on that, we will show that $\|W_n(\theta)\| = o_p(\sqrt{n}(\delta \wedge \|\theta - \theta_0\|))$ uniformly on $\theta \in \widetilde{K}^c$, and therefore

$$\|W_n(\theta)\| = o_p(\|E[S_n(\theta)]\|), \text{ uniformly in } \theta \in \tilde{K}^c. \qquad (A.46)$$

For the case that $\delta \leq \|\theta - \theta_0\| \leq T_n$ it suffices to have $\sqrt{d \log n}T_n^\alpha + n^{-1/2}d^{3/2}\log n = o(n^{1/2})$. On the other hand, for $C\sqrt{d/n} \leq \|\theta - \theta_0\| \leq \delta$ it suffices to have $\sqrt{d \log n}\|\theta - \theta_0\|^\alpha + n^{-1/2}d^{3/2}\log n = o(\sqrt{n}\|\theta - \theta_0\|)$. Indeed, $\sqrt{d \log n}\|\theta - \theta_0\|^\alpha = (\sqrt{n}\|\theta - \theta_0\|)$ if $\sqrt{d \log n} = o(\sqrt{n}\|\theta - \theta_0\|^{1-\alpha})$, which is implied by $\sqrt{d \log n} = o(\sqrt{n}(d/n)^{\frac{1-\alpha}{2}})$. Moreover, $n^{-1/2}d^{3/2}\log n = o(\sqrt{n}\|\theta - \theta_0\|)$ if $n^{-1/2}d^{3/2}\log n = o(\sqrt{n}\sqrt{d/n})$. All of the above conditions hold under condition ZE.3.

Step 3. Lower bound on $\|S_n(\theta)\|$. We will show that

$$\|S_n(\theta)\|^2 = \|E[S_n(\theta)] + S_n(\theta_0) + W_n(\theta)\|^2 \geq \frac{1}{2}\|E[S_n(\theta)]\|^2 \qquad (A.47)$$

uniformly for all $\theta \in \widetilde{K}^c$ wp $1 - 2\varepsilon$.

For any two vectors a and b, we have $\|a + b\|^2 \geq (\|a\| - \|b\|)^2 = \|a\|^2 - 2\|a\|\|b\| + \|b\|^2 \geq \|a\|^2(1 - 2\|b\|/\|a\|)$. Applying this relation with $a = E[S_n(\theta)]$ and $b = W_n(\theta) + S_n(\theta_0)$,(A.45), and (A.46), we obtain (A.47).

Step 4. Bounding the integrals. Using (A.47) and ZE.2 wp $1 - 3\varepsilon$

$$
\begin{aligned}
\int_{\widehat{K}^c} \exp(-\|S_n(\theta)\|^2)d\theta &\leq \int_{\widetilde{K}^c} \exp(-\|S_n(\theta)\|^2)d\theta \\
&\leq \int_{\widetilde{K}^c} \exp(-\tfrac{1}{2}\|E[S_n(\theta)]\|^2)d\theta \\
&\leq \int_{\widetilde{K}^c} \exp(-\tfrac{1}{2}\mu n\|\theta - \theta_0\|^2)d\theta + \int_{\widetilde{K}^c} \exp(-\tfrac{1}{2}\mu n\delta^2)d\theta \\
&\leq (2\pi)^{d/2}(n\mu)^{-\frac{d}{2}}P(\|U\| > \tilde{C}\sqrt{d/n}) + \exp(-\tfrac{1}{2}\mu n\delta^2)\text{vol}(\Theta_n) \\
&\leq (2\pi)^{d/2}(n\mu)^{-\frac{d}{2}}\exp\left(-\frac{(\tilde{C}-1/\sqrt{\mu})^2\mu}{2}d\right) + \nu_d T_n^d \exp(-\tfrac{1}{2}\mu n\delta^2)
\end{aligned}
$$

where ν_d is the volume of the d-dimensional unit ball, which goes to zero as d grows, and $U \sim N(0, \frac{1}{\mu n}I_d)$. In the first line we used the inclusion (A.44), and in the last line we used a standard Gaussian concentration inequality, Proposition 2.2 in Talagrand [51], and the fact that $E[\|U\|] \leq (E[\|U\|^2])^{1/2} = \frac{1}{\sqrt{\mu}}\sqrt{d/n}$.

On the other hand, by Lemma 7 we have

$$-\|S_n(\theta)\|^2 + \|S_n(\tilde{\theta})\|^2 = n\|A(\theta - \tilde{\theta})\|^2 + o_p(1)$$

uniformly for $\theta \in \widehat{K}$. This yields that wp $1 - \varepsilon$

$$
\begin{aligned}
\int_{\widehat{K}} \exp(-\|S_n(\theta)\|^2) \ d\theta &\geq \exp(-\|S_n(\tilde{\theta})\|^2)\int_{\widehat{K}} \exp(-n\|A(\theta - \tilde{\theta})\|^2 + o_p(1))d\theta \\
&\geq \exp(-C_2 d)\int_{\widehat{K}} \exp(-C_1 n\|\theta - \tilde{\theta}\|^2)d\theta \\
&\geq \exp(-C_2 d)(2\pi)^{\frac{d}{2}}(C_1 n)^{-\frac{d}{2}}(1 - P(\|U\| \leq C\sqrt{d/n})) \\
&\geq \exp(-C_2 d)(2\pi)^{\frac{d}{2}}(C_1 n)^{-\frac{d}{2}}(1 - o(1))
\end{aligned}
$$

where constant C_1 is maximal eigenvalue of $A'A$, constant C_2 is such that $\|S_n(\tilde{\theta})\|^2 \leq C_2 d$ wp $1-\varepsilon$ by Lemma 7, $U \sim N(0, \frac{1}{C_1 n} I_d)$. In the last line we used the standard Gaussian concentration inequality, Proposition 2.2 in Talagrand [51], with constant $C > 2/\sqrt{C_1}$ to get $P(\|U\| \leq C\sqrt{d/n}) = o(1)$.

Finally, we obtain that wp $1 - 5\varepsilon$

$$\frac{\int_{\widehat{K}^c} \exp(-\|S_n(\theta)\|^2) d\theta}{\int_{\widehat{K}} \exp(-\|S_n(\theta)\|^2) d\theta} \leq \frac{(2\pi)^{\frac{d}{2}} (\mu n)^{-\frac{d}{2}} \exp\left(-\frac{(\tilde{C}-1/\sqrt{\mu})^2 \mu}{2} d\right) + \nu_d T_n^d \exp(-\frac{1}{2}\mu n \delta^2)}{\exp(-C_2 d) \, (2\pi)^{d/2} (C_1 n)^{-d/2} (1 + o(1))}$$

where the right hand side is $o(1)$ by choosing $\tilde{C} > 0$ sufficiently large, and noting that terms $(2\pi)^{d/2} n^{-d/2}$ cancel and that $d \ln T_n = o(n)$ by condition ZE.3.

Since $\varepsilon > 0$ can be set as small as we like, the conclusion follows. ∎

Appendix B: Bounding log-β-concave functions

Lemma 9 *Let $f : \mathbb{R} \to \mathbb{R}$ be a unidimensional log-β-concave function. Then there exists a logconcave function $g : \mathbb{R} \to \mathbb{R}$ such that*

$$\beta g(x) \leq f(x) \leq g(x) \ \ for \ every \ \ x \in \mathbb{R}.$$

Proof. Consider $h(x) = \ln f(x)$ a $(\ln \beta)$-concave function. Now, let m be the smallest concave function greater than $h(x)$ for every x, that is,

$$m(x) \ \ = \ \ \sup \left\{ \sum_{i=1}^{k} \lambda_i h(y_i) : k \in \mathbf{N}, \lambda \in \mathbb{R}^k, \lambda \geq 0, \sum_{i=1}^{k} \lambda_i = 1, \sum_{i=1}^{k} \lambda_i y_i = x \right\}.$$

Recall that the epigraph of a function w is defined as $epi_w = \{(x, t) : t \leq w(x)\}$. Using our definitions, we have that $epi_m = \text{conv}(epi_h)$ (the convex hull of epi_h), where both sets lie in $\mathbb{R}^2$. In fact, the values of m are defined only by points in the boundary of $\text{conv}(epi_h)$. Consider $(x, m(x)) \in epi_m$, since the epigraph is convex and this point is on the boundary, there exists a supporting hyperplane H at $(x, m(x))$. Moreover, $(x, m(x)) \in \text{conv}(epi_h \cap H)$. Since H is one dimensional, $(x, m(x))$ can be written as convex combination of at most 2 points of epi_h.

Furthermore, by definition of log-β-concavity, we have that

$$\ln 1/\beta \geq \sup_{\lambda \in [0,1], y, z} \lambda h(y) + (1 - \lambda) h(z) - h(\lambda y + (1 - \lambda) z).$$

Thus, $h(x) \leq m(x) \leq h(x) + \ln(1/\beta)$. Exponentiating gives $f(x) \leq g(x) \leq \frac{1}{\beta} f(x)$, where $g(x) = e^{m(x)}$ is a logconcave function. ∎

References

[1] Y. F. Atchade (2006). An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift, Methodol. and Comput. in Applied Probab., 8: 235–254.

[2] D. Applegate and R. Kannan (1993). Sampling and Integration of Near Logconcave Functions, Proceedings 23th ACM STOC, 156–163.

[3] O. Barndorff-Nielsen (1978). Information and exponential families in statistical theory. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester.

[4] A. Belloni and V. Chernozhukov (2007). Posterior inference in curved exponential families under increasing dimensions, MIT and Duke Technical Report.

[5] A. Belloni and V. Chernozhukov (2008). Conditional Quantile Processes under Increasing Dimension, MIT and Duke Technical Report.

[6] P. J. Bickel and J. A. Yahav (1969). Some contributions to the asymptotic theory of Bayes solutions, Z. Wahrsch. Verw. Geb 11, 257–276.

[7] D. Blackwell (1985). Approximate normality of large products, Technical Report 54, Department of Statistics, Univ. Calif. Berkeley.

[8] Bunke, O., Milhaud, X. (1998) Asymptotic behavior of Bayes estimates under possibly incorrect models. The Annals of Statistics 26 (2), 617–644.

[9] G. Casella and C. P. Robert (1999). Monte Carlo Statistical Methods, Springer Texts in Statistics.

[10] G. Chamberlain (1987). Asymptotic efficiency in estimation with conditional moment restrictions, Journal of Econometrics 34, no. 3, 305–334

[11] V. Chernozhukov and H. Hong (2003). An MCMC approach to classical estimation, Journal of Econometrics 115 293-346.

[12] S. Chib (2001). Markov Chain Monte Carlo Methods: Computation and Inference, Handbook of Econometrics, Volume 5, by J.J. Heckman and E. Leamer, Elsvier Science.

[13] S. G. Donald, G. W. Imbens, W. K. Newey (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. Journal of Econometrics, 117, no. 1, 55–93.

[14] B. Efron (1978). The geometry of exponential families, Annals of Statistics, 6, no. 2, 362–376.

[15] G. S. Fishman (1994). Choosing sample path length and number of sample paths when starting at steady state, Operations Research Letters, Vol. 16, No. 4, November, pp. 209–220.

[16] A. Frieze, R. Kannan and N. Polson (1994). Sampling from log-concave functions, Annals of Applied Probability 4, pp. 812-834.

[17] A. Gelman, G. O. Roberts and W. R. Gilks (1996). Efficient Metropolis jumping rules, Bayesian Statistics V, ed. by J. M. Bernardo et al., 599–607. Oxford University Press.

[18] J. Geweke and M. Keane (2001). Computationally Intensive Methods for Integration in Econometrics, Handbook of Econometrics, Volume 5, by J.J. Heckman and E. Leamer, Elsvier Science.

[19] C. J. Geyer (1992). Practical Markov Chain Monte Carlo, Statistical Science, Vol. 7, No. 4, 473–511.

[20] S. Ghosal (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity, Journal of Multivariate Analysis, vol 73, 49–68.

[21] L. P. Hansen and K. J. Singleton (1982). Generalized instrumental variables estimation of nonlinear rational expectations models, Econometrica 50, no. 5, 1269–1286.

[22] X. He and Q.-M. Shao (2000). On Parameters of Increasing Dimenions, Journal of Multivariate Analysis, vol 73, 120–135.

[23] P. J. Huber (1993). Robust regression: asymptotics, conjectures and Monte Carlo. Ann. Statist. 1 (1973), 799–821.

[24] I. Ibragimov and R. Has'minskii (1981). Statistical Estimation: Asymptotic Theory, Springer, Berlin.

[25] G. W. Imbens (1997). One-step estimators for over-identified generalized method of moments models. Review of Economics Studies 64, no. 3, 359–383.

[26] M. Jerrum and A. Sinclair (1988). Conductance and the rapid mixing property for Markov chains: the approximation of permanent resolved, Proceedings of the 20th Annual ACM symposium on Theory of Computing, pp. 235-244.

[27] M. Jerrum and A. Sinclair (1989). Approximating the permanent, SIAM Journal on Computing, 18:6, 1149–1178.

[28] R. Kannan and G. Li (1996). Sampling according to the multivariate normal density, 37th Annual Symposium on Foundations of Computer Science (FOCS '96), pp. 204.

[29] R. Kannan, L. Lovász, and M. Simonovits (1995). Isoperimetric Problems for Convex Bodies and a Localization Lemma, J. Discr. Comput. Goem., volume 13, pp. 541–559.

[30] R. Kannan, L. Lovász, and M. Simonovits (1997). Random walks and an $O^*(n^5)$ volume algorithm for convex bodies, Random Structures and Algorithms, volume 11, pp. 1–50.

[31] C. Kipnis and S. R. S. Varadhan (1988). Central limit theorem for additive functionals of reversible processes and applications to simple exclusions, Comm. Math. Phys., 104, 1-19.

[32] R. Koenker (1988). Asymptotic theory and econometric practice, Journal of Applied Econometrics 3, 139-147.

[33] E. L. Lehmann and G. Casella (1998). Theory of point estimation. Second edition. Springer Texts in Statistics. Springer-Verlag, New York.

[34] J. S. Liu, L. Tian and L.J. Wei. Implementation of estimating-function based inference procedures with MCMC samplers. Journal of American Statistical Association, to appear.

[35] J. S. Liu (2001). Monte Carlo Strategies in Scientific Computing, Springer-Verlag, New York.

[36] L. Lovász (1999). Hit-and-run mixes fast, Mathematical Programming, Ser. A 86: 443–461.

[37] L. Lovász and M. Simonovits (1993). Random Walks in Convex Bodies and an Improved Volume Algorithm, Random Structures and Algorithms, 4:4, 359–412.

[38] L. Lovász and S. Vempala (2007). The Geometry of Logconcave Functions and Sampling Algorithms, Random Structures and Algorithms, 30:3, Pages

307 - 358.

[39] L. Lovász and S. Vempala (2003). Hit-and-Run is Fast and Fun, Technical report MSR-TR-2003-05.

[40] L. Lovász and S. Vempala (2006). Hit-and-Run from a Corner, SIAM Journal on Computing, v.35 n.4, p.985-1005.

[41] N. Polson (1996). Convergence of Markov Chain Monte Carlo Algorithms, Bayesian Statistics 5 (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press.

[42] S. Portnoy (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. Annals of Statistics 16, no. 1, 356–366.

[43] G. O. Roberts, A. Gelman and W. R. Gilks (1997). Weak convergence and optimal scaling of random walk metropolis algorithms, The Annals of Applied Probability, Vol. 7, No. 1, 110–120.

[44] G. O. Roberts and J. S. Rosenthal (2001). Optimal Scaling for Various Metropolis-Hastings Algorithms Statistical Science, Vol. 16, No. 4, pp. 351–367.

[45] G. O. Roberts and R. L. Tweedie (1996). Exponential Convergence of Langevin Diffusions and Their Discrete Approximations, Bernoulli, 2, pp. 341-364.

[46] X. Shen (2002). Asymptotic Normality of Semiparametric and Nonparametric Posterior Distributions, Journal of the American Statistical Association, Vol. 97, No. 457, 222–235.

[47] C. J. Stone, M. H. Hansen, C. Kooperberg, and Y. K. Truong (1997). Polynomial splines and their tensor products in extended linear modeling. With discussion and a rejoinder by the authors and Jianhua Z. Huang. Annals of Statistics, 25, no. 4, 1371–1470.

[48] O. Stramer and R.L. Tweedie (1999). Langevin-Type Models I: Diffusions with Given Stationary Distributions and their Discretizations, Methodology and Computing in Applied Probability, Vol. 1, Number 3, 283–306.

[49] A. W. van der Vaart and J. A. Wellner (1996). Weak Convergence and Empirical Processes Spring Series in Statistics.

[50] S. Vempala (2005). Geometric Random Walks: A Survey, Combinatorial and Computational Geometry, MSRI Publications Volume 52.

[51] M. Talagrand (1994). Sharper bounds for gaussian and empirical processes, The Annals of Probability, Vol. 22, No. 1, 28-76.